Solar Energy 259 (2023) 257–276

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Soiling determination for parabolic trough collectors based on operational data analysis and machine learning

Alex Brenner [a,b,*], James Kahn [c,d], Tobias Hirsch [a], Marc Röger [e], Robert Pitz-Paal [f,b]

[a] German Aerospace Center (DLR), Institute of Solar Research, Wankelstrasse 5, 70563 Stuttgart, Germany
[b] RWTH Aachen University, Chair of Solar Technology, Germany
[c] Helmholtz AI, Germany
[d] Karlsruhe Institute of Technology (KIT), Steinbuch Centre for Computing, Hermann-von-Helmholtz Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
[e] German Aerospace Center (DLR), Institute of Solar Research, Paseo de Almería 73, E-04001 Almería, Spain
[f] German Aerospace Center (DLR), Institute of Solar Research, Linder Höhe, 51147 Cologne, Germany

ARTICLE INFO

Keywords:
Parabolic trough
Soiling
Machine learning
Artificial neural network
Concentrated solar power

ABSTRACT

Advanced cleaning strategies for parabolic trough collectors at concentrated solar power plants maximize the yield and minimize the costs for cleaning activities. However, they require information about the current soiling level of each collector. In this work, a novel, data-driven method for soiling estimation with machine learning for parabolic trough collectors is developed using gloss values as a surrogate for soiling values. Operational data and meteorological data from the solar field Andasol-3 with changing time horizons are used together with various Machine Learning techniques to estimate the soiling of every collector in the field. The best results were achieved with a Decision Tree model, with a coefficient of determination of $R^2 = 0.77$ from the maximum value of 1 and a mean squared error of $MSE = 6.14$ for the determination of specific soiling values. A second metric to evaluate the quality of soiling predictions from the models classifies whether soiling is above or below a cleaning threshold was also investigated. Model results are compared to soiling measurements that indicate the need for cleanings. Cleaning recommendations are derived and compared with the current fixed-time cleaning schedule of Andasol-3. All models show an improvement over the cleaning schedule currently in use. The use of a Decision Tree model increases the detected necessary cleanings by 12.2 %, while the number of unnecessary cleanings are reduced by 14.3 %. This has the potential to reduce operational costs and increase the solar field yield. The dataset used in this work is made publicly available https://doi.org/10.5281/zenodo.7061913, along with the code to reproduce all results, which can be found at https://doi.org/10.5281/zenodo.7554806.

1. Introduction

An established source of renewable energy with the capability to deliver dispatchable electricity is concentrated solar power (CSP). CSP plants use direct solar irradiation, transforming it into thermal energy. The thermal energy can then be used as process heat or to run a power cycle and produce electricity directly. Most CSP sites are located in the sun belt region with high direct normal irradiance (DNI). However, these regions often have arid climates and a high dust load potential. High dust loads may lead to dust deposition on the CSP plant mirrors, an effect known as soiling. Soiling is a major source of performance loss in CSP solar fields, with 3%–4% reduction of solar power production, causing annual revenue losses of 3 to 5 billion € [1]. In comparison to photovoltaic (PV) systems, the losses due to soiling in CSP systems can be 8 to 14 times higher [2]. In general, CSP systems are directly exposed to the harsh environmental conditions and therefore soiling is always present. Since cleaning is costly, has a high water consumption, and speeds up the degradation of the mirrors, it should be reduced to a minimum. Yet, an efficient cleaning strategy can only be executed if the soiling level of each mirror is accurately known.

Currently, handheld devices are used to record soiling levels in the collector field. These devices deliver accurate readings, but only spot measurements and the procedure is very time consuming. In this work, a data-driven model for the estimation of soiling of each collector in the parabolic trough solar field is presented. The model only requires operational data that is already recorded in the power plant or data which is readily available from local meteorological stations and services. Our approach uses the spatially distributed measurement instrumentation already installed in the parabolic trough field (see [3]). We create soiling specific model inputs, so-called features, from the measurement data and evaluate their usefulness for the task. With these

* Corresponding author at: German Aerospace Center (DLR), Institute of Solar Research, Wankelstrasse 5, 70563 Stuttgart, Germany.
E-mail addresses: alex.brenner@dlr.de (A. Brenner), james.kahn@kit.edu (J. Kahn).

https://doi.org/10.1016/j.solener.2023.05.008
Received 19 September 2022; Received in revised form 25 April 2023; Accepted 5 May 2023
Available online 24 May 2023

specific features we compare different models and datasets in order to check whether a small dataset and a simple model delivers sufficient accuracy, or if a larger dataset and a more complex model are required.

The novelty of this approach is the use of already available data collected at the power plant and the ability to continuously deliver soiling information for each collector in the field. Therefore, implementing such a soiling determination system is expected to have low operation and maintenance costs, since it only requires software and no additional hardware or personnel costs are expected. Operational and meteorological data are continuously measured at the power plant and therefore available for soiling determination. Soiling can thus be determined steadily for every collector in the solar field. This spatially resolved soiling information is especially interesting for individual cleaning operations or performance estimations. Determination of soiling from operational and meteorological data also has the potential of replacing soiling field measurements and as a consequence, the costs for monitoring the solar field soiling level can be reduced. For this use-case the model needs to be further adjusted to enable online use. Due to limited data availability, this has not yet been implemented in the model presented in this work.

The original contributions of this work are:

- First usage of operational and meteorological data to create a data-driven soiling model for parabolic trough fields.
- Comparison of various machine learning models and dataset sizes for the task of soiling determination.
- Evaluation of model performance for the application of soiling prediction and provision of cleaning recommendations.
- Investigation of the most important model inputs for soiling prediction via permutation feature importance.
- The code to reproduce this work's results is made publicly available at https://doi.org/10.5281/zenodo.7554806, along with the dataset used https://doi.org/10.5281/zenodo.7061913.

Section 2 first introduces all datasets and models used in the work. Section 3 describes the metrics used for the evaluation of the model results. In Section 4 the model results are first evaluated with a direct comparison to reference gloss measurements and then as a classification problem. Section 5 contains the conclusion and an outlook on future research.

1.1. Related work

In PV research, Pulipaka et al. [4] modeled the soiling loss with a neural network and linear regression. Model inputs were the percentage of different particle sizes and the incident horizontal irradiance. In contrast to our work, it was an experimental setup with artificial soiling used as training data. In most applications the exact particle size distribution is not known and may also change due to different weather conditions. Javed et al. [5] use natural soiling data together with environmental data from a PV field in Doha, Qatar to estimate the change in cleanliness. They achieved the best results with a simple Neural Network and compared it to a Linear Regression approach. Nevertheless, just using environmental data limited the best R^2 value to 0.537 for the Neural Network approach. Another experimental setup was used by Laarabi et al. [6] in order to gather soiling and environmental data for the model training. They use only the PV glass cover for this setup and collected data from more than eight months, with daily measurements. Their Neural Network model predicts the soiling rate with an R^2 value of 0.928. Still, the used dataset was small with regular measurements every day at noon from April 20 to December 31, 2016. The model is expected to have low generalization capabilities, due to the dataset not covering an entire year. Adapting the model to new sites would require an additional experimental setup at the new location to perform further data collection. Chiteka et al. [7] also compared a Neural Network and a Linear Regression approach with environmental data, achieving comparable results. Simal Pérez et al. [8] added the short circuit current and solar altitude to the environmental dataset, showing an improvement in predictive performance. Since the focus of the latter three publications is on non-tracking PV systems with fixed tilt angles, it is difficult to apply the models to parabolic troughs. Besides the PV models using machine learning techniques, a simple physical soiling model was developed by Coello and Boyle [9]. They applied their model on seven different locations and used particulate matter (PM) values as a measure of dust concentration in the air (PM_{10}, $PM_{2.5}$), tilt, tracking, and rain data to determine the efficiency decrease due to soiling. The physical model is applicable on different locations without collecting large datasets and is implemented in `pvlib python` [10]. In summary, there are several approaches to determine soiling for PV modules, but they are not directly applicable to parabolic trough fields as they either do not deliver spatially resolved soiling estimates, are mostly not applicable to tracked systems, or do not use the operational data of the solar field.

Publications about soiling determination for CSP systems mainly focus on the development of physical models in order to be site independent [11,12]. Both of these models use meteorological data as inputs and are not specialized for the use case of parabolic trough fields. Sbarbaro et al. [13] used simulation data from a parabolic trough model in order to estimate the soiling level from collector temperature data. However, a validation with real measurement data was not performed. Conceição et al. [14] used meteorological data as model input and data from the Tracking Cleanliness Sensor (TraCS) System [15] as target values. One drawback is that this approach does not take into account the spatial influence of soiling within a solar field. El Gallassi et al. [16] also used the TraCS system to record soiling target values. Based on this, they frame the problem as a multi-class classification problem, where each class corresponds to a different soiling level. Due to limited data, they are unable provide test examples for all classes, meaning the generalization capability of their produced model is unverified.

These soiling models from CSP applications are not directly usable in the application of parabolic trough fields since they do not deliver spatially resolved soiling information, are not validated for real data from the solar field, or require additional measurement instrumentation. None of them use the spatially distributed measurement instrumentation, which is already available in parabolic trough fields. Beside the shown modeling approaches there are also advanced measuring approaches, which deliver spatially distributed measurements. Wolfertstetter et al. [17] showed a promising application of an airborne system for soiling measurements. This system might be a good supplement to a model based implementation, which can continuously determine the soiling level with low implementation and operational costs, but with higher uncertainties compared to a pure measurement approach.

The publications highlighted above already show promising results using models to estimate the soiling level either for PV or CSP. The spatially distributed measurement instrumentation in parabolic trough plants can be especially useful for delivering raw data for a data-driven model to estimate soiling individually for each collector in the solar field.

2. Datasets and models

2.1. Available raw data

The data used in this work is taken from the Andasol-3 power plant in Spain and cover the years 2015 to 2017. It is measured at the parabolic trough collectors, at the meteorological stations in the solar field, at additional meteorological stations near the power plant, or is publicly available from weather services. Parabolic trough fields are usually equipped with numerous measurement instruments at each collector. A detailed review of the typical instrumentation is given by Brenner et al. [3]. For the purpose of developing soiling

Table 1
Example of dataset with different records in rows, labels which should be predicted in the first column, and features in the other columns.

Record	Labels	Features			
date:collector	gloss	week of year sin	week of year cos	...	pm10 avg
2017.03.01:LG28	83.55	0.89	0.46	...	2.87e−8
2015.09.28:RF11	91.25	−0.99	0.12	...	9.64e−9
...	...	...	...	...	...

determination models we make use of the instrumentation at the collector level, the meteorological instrumentation, and additional gloss measurements. The raw data from the power plant has a temporal resolution of one to ten seconds. In this work, gloss measurements collected with the Zehntner ZGM 1110 are used to generate an approximation of the soiling level in the solar field. The gloss measurements in Andasol-3 are usually carried out in the evenings, roughly every two days at 40 predefined locations for each subfield. Two measurements are taken at every location at roughly the same time. At the collector, gloss measurements are taken at the central mirror facet at the second south or north solar collector element (SCE) as seen from the central drive pylon. The southern or northern solar collector element is selected depending on which is closer to the road in the solar field. In total 18,540 gloss measurements from 2015 to 2017 are available. The gloss meter[1] has a value range between 0 (completely soiled) and 100 (clean mirror), and is calibrated with a clean mirror sample. Small deviations in the calibration procedure can lead to gloss values slightly above 100 gloss units. A histogram of the gloss measurements is shown in Fig. B.20 in the appendix. Reflectance measurements would be preferable since glossmeter measurements are not recommended for evaluating specular reflectance [19]. However, the frequent use of gloss measurements as an indicator of soiling at the Andasol-3 power plant shows that they are used in industry and are therefore available in sufficient numbers to develop a soiling model. Furthermore, Brooks and Schwar [20] showed that gloss measurements have an approximately linear relationship with the percentage area covered by dust. Fig. 1 shows a top view of a parabolic trough field. The position of each collector in the field is given by the x-axis, i.e. from west to east, and y-axis, i.e. from south to north. Corresponding wind components are indicated by u and v. Wind component u is positive for wind from west to east, and therefore perpendicular to the collector axis, and wind component v is positive for wind from south to north and aligned with the collector axis. The measurement instrumentation is positioned at the center of each collector. The meteorological stations are positioned at each corner of the solar field and in the middle of the solar field.

2.2. Features and datasets

Datasets are tabular numerical data with features as columns and coherent records as rows (see Table 1). The feature collector position x is for example an integer value created from the raw collector name string. The created features, with their usage in the different datasets, are summarized in the appendix in Table B.7 and their distribution is shown in the histograms in Figs. B.18 and B.19. The number of records is limited by the amount of available gloss measurements. These measurements serve as labels for model training and evaluation. Since gloss measurements are available at most once per day, the features created from solar field data and meteorological time-series are reduced to one value per gloss measurement for each feature. This is done with averaged, maximum, minimum, or temporal values according to the feature definition.

The created features can be categorized into three groups. The first only takes into account the measurement data from the same day on which the gloss values are evaluated. We assume that a certain soiling level is also recognizable in the operational data for this day. This means that we can create features of this group only on days where gloss measurements are available. Features from this category are the: time of year, daily average irradiance on collector aperture, daily cumulative irradiance, focus factor[2] of the collector, dumping value[3] of the collector, and collector temperature. Changing environmental conditions during the year due to the changing seasons strongly influence soiling. The time of year feature encapsulates this seasonal influence. It is represented by the week of the year and in order to only include the seasonal aspect is split into two features (sine and cosine). This changes the continuous numeric feature to a cyclic one. The approach for encoding time is explained by London [22] and Vaswani et al. [23] use a similar method that is intended to help the model understand the relative position of an element in a sequence.

The second feature group contains positional information about the collector. The position of the collector in the field is given by an x- and y-position and the position within the loop according to Fig. 1. The position of the collector in the field should help estimate inhomogeneous soiling from the cooling tower, as stated by Cohen et al. [24], or shielding effects from other collectors.

The third feature group includes features derived from the time span between the last cleaning of the collector until the gloss measurement is taken. The first feature of this group is the days since the last cleaning of the collector. We consider this important as soiling accumulates over time. It is calculated from a command which initiates the positioning of the collectors for the cleaning procedure. Wind influence is equally considered to be of importance because at a certain wind speed dust can be lifted up from the ground and settle on the mirrors. The feature is included as averaged and thresholded wind values. For the averages, the wind speed vector is expressed using two wind components u and v. For the thresholds, the magnitude of wind speed without the split in two components is used. The threshold is defined by the minimal wind speed at which the first dust particles are lifted from the ground. This velocity is called minimal saltation fluid threshold u^*_{ft}. In order to use the wind measurement from the meteorological station, which is measured at 10 m above the ground, the logarithmic law of the wall is used to extrapolate the wind speed at ground level (see Eq. (1)), where $\kappa \approx 0.40$ is the von Kármán's constant, and $z_0 = 3.33 \times 10^{-4}$ m is the surface roughness. In our case $u^*_{ft} = 0.23$ m/s, determined from [25], results in a mean horizontal fluid velocity of $\overline{U}_x(z) = 5.9$ m/s at a measurement height of $z = 10$ m.

$$\overline{U}_x(z) = \frac{u^*_{ft}}{\kappa} * \ln(\frac{z}{z_0}) \tag{1}$$

With this definition, the total hours above this threshold since last cleaning, the hours since the last time this threshold was reached, and the average collector angle during the time above the threshold are used as features. Dew formation is another meteorological influence which may lead to higher soiling or have a cleaning effect, as discussed by Conceição et al. [14],Caron and Littmann [26] and Mehos et al.

[1] Zehntner ZGM 1110, using 20° measurement angle between incident light and perpendicular with repeatability of 0.1 gloss units and reproducibility of 0.5 gloss units [18].

[2] Result of the acceptance angle curve using the deviation of the collector tracking angle and the sun angle [21].

[3] Percentage value of power curtailment for a certain collector, given by solar field control system.

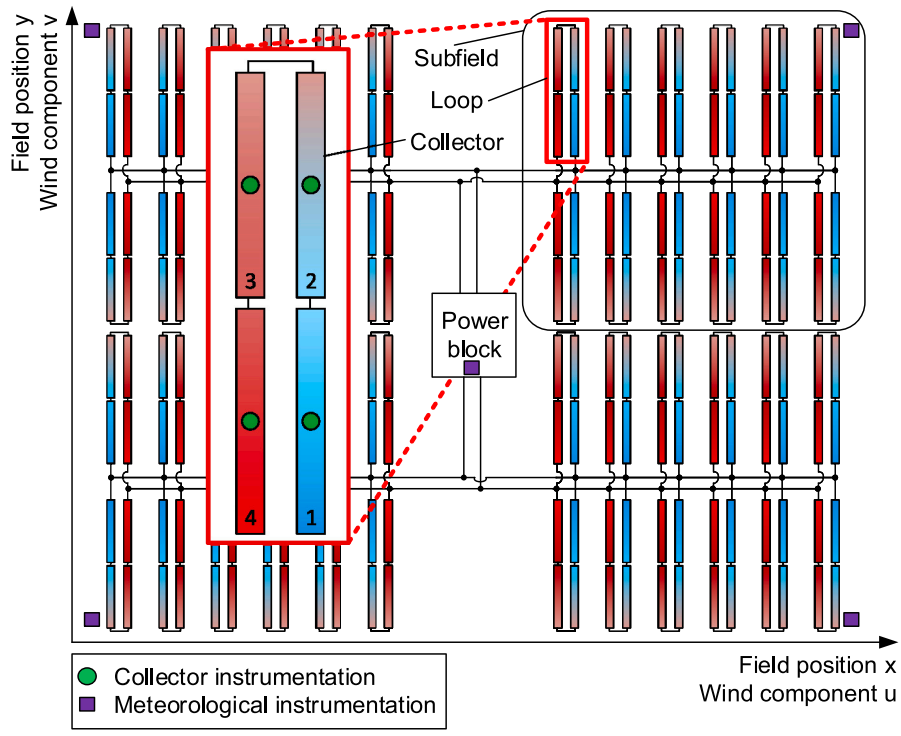

Fig. 1. Layout of parabolic trough field (not to scale, for illustration only) including field position (x, y), collector position in loop (1, 2, 3, 4), wind components (u, v), and positions of measurement and meteorological instrumentation.

[27]. The accurate measurement of dew formation would require an additional dew sensor or at least a mirror temperature sensor. Both of them are not available, therefore a simple approximation of dew formation is given by the difference between dew point temperature[4] and ambient air temperature. Included features comprise dew formation average values and threshold values. To estimate times with dew formation, a threshold defined as $T_{amb} - T_{dew} = 5\,\mathrm{K}$ is set. If the difference between dew point temperature and ambient air temperature is below this threshold, dew formation is expected. Similar to the wind speed threshold, the hours under the dew point threshold since last cleaning, hours since the threshold was last reached, and the average collector angle during this period are included as features. Natural cleaning effects by precipitation are included with an average value, maximum value, and threshold values. A threshold for a natural cleaning effect of a rainfall event has already been discussed in literature and ranges from 0.3 mm [29] up to 10 mm [30], or is assumed to be not applicable [31]. We included a theoretical cleaning threshold of 5 mm, and additionally included features for dry periods, since they showed good correlation with soiling ratio for PV soiling [29]. The thresholding approach for precipitation features is comparable to that used for wind speed. Feature are hours above precipitation threshold since last cleaning, hours since the threshold is reached the last time, and average collector angle during this period. Features for the dry period are hours of the longest dry period with 0 mm precipitation since last cleaning, hours since the last time with a dry period, and average collector angle during the dry period. The last feature considered was an average PM_{10} value[5] from a reanalysis dataset with $0.75° \times 0.75°$ horizontal resolution. PM_{10} values show high correlations to soiling rates, and therefore appear to be the most relevant PM value [32]. A selection of the data used in the datasets is shown in Figs. A.11, A.12, A.13, A.14, A.15, A.16 and A.17. The data not shown here cannot be presented in a meaningful way over a uniform time axis due to their dependence on individual cleaning times of the collectors.

[4] Calculated from relative humidity and air temperature according to the formulas given in [28].

[5] Generated using Copernicus Atmosphere Monitoring Service Information [2021]

2.2.1. Dataset preparation

We divide the dataset into three distinct subsets: training (80 %), validation (10 %), and test (10 %). The largest subset is used to train the model (training data). In this subset, the measured gloss values are used to fit the model to the data. A smaller part of the dataset is used for validating the trained model (validation data). The model uses the validation data without the measured gloss values and predicts the gloss instead. This is then used to estimate the performance and tune the hyper-parameters of the model. The last fraction of the dataset is used for the final evaluation of the model performance on unseen data (test data). In order to use each data record independently from the other records, we randomly shuffled the entire dataset to avoid data patterns which may disturb the model prediction. Therefore, the information about the chronological sequence of the measurements is removed. This is possible because the original time-series data was first converted to tabular data, as described above. The datasets used for all models except for Decision Trees were scaled using z-score normalization for each feature independently, defined as

$$z_i = \frac{x_i - \mu_i}{\sigma_i},$$

for feature i with value x_i, mean μ_i and standard deviation σ_i. The `StandardScaler` class from Scikit-learn library, version 0.23.1, was used for this purpose [33].

2.2.2. Train, validation and test distribution

In order to ensure the dataset subsets sufficiently represent the true underlying data distribution, we use a Kolmogorov–Smirnov test [34] to verify the equality of the feature distributions between the training data and the validation and test data. The Kolmogorov–Smirnov test calculates the distance between two distributions. If it is close to zero it is likely that the two distributions are identical. The null hypothesis

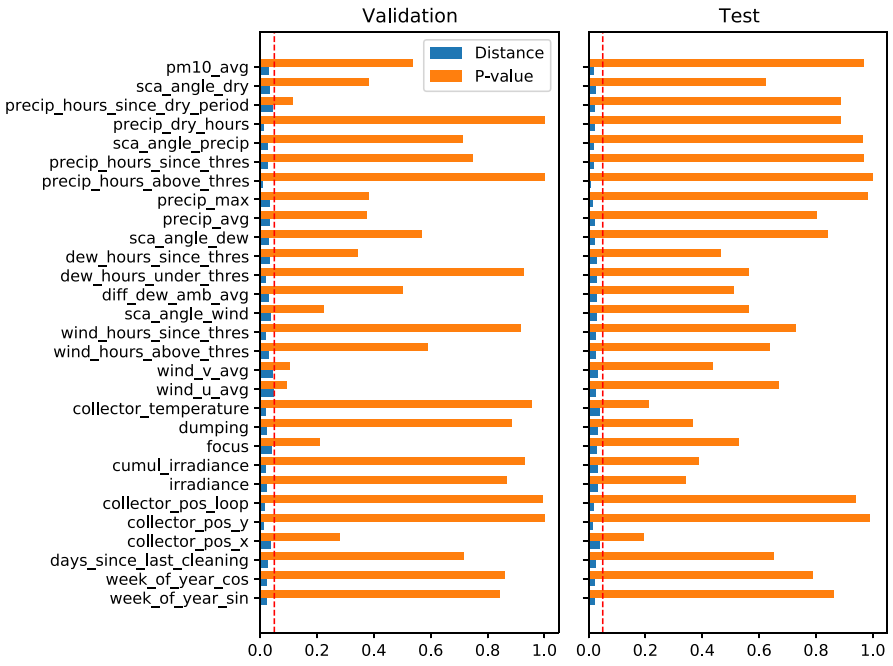

Fig. 2. Kolmogorov–Smirnov test comparing training dataset with validation and test dataset. Both subsets show small distance values or high p-values or both, which shows the similarity of the data distributions. All p-values are above the chosen significance threshold of 0.05 (marked with red dashed line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is that the two feature distributions, training data compared with validation and test data, have identical distributions. We also calculate the corresponding p-value for each feature. A high p-value above the significance level makes it likely that the two distributions are identical and therefore the null hypothesis would be correct. The results are collected in Fig. 2 and show small distance values and high p-values, above the commonly used threshold of 0.05, for all features. As long as the p-values are above the significance level, deviations between validation and test are tolerable. The results from this statistical test can be further visualized with a histogram plot of the features in the appendix in Figs. B.18 and B.19, where comparable distributions for all features can be seen. With this test we can assume that a model trained with the training data subset delivers comparable predictions with validation and test data.

2.3. Classical machine learning models

We investigate four common classical machine learning models implemented in Scikit-learn library [33], version 0.23.1, which either have a high interpretability or deliver an uncertainty estimation. These models are briefly described in the following and serve as a baseline to be compared with a neural network approach described in Section 2.4. Configuration parameters of the presented models are given in the appendix in Table B.8.

Linear Regression The simplest model we used is a Linear Regression (LR) model, which can infer a linear relationship between the desired output and the dependent variables. It is fitted with ordinary least squares method.

Support Vector Regression Support Vector Regression (SVR) is an extension of Support Vector Classification and widely used in machine learning applications. By using the radial basis function kernel for SVR the model is able to handle non-linearities.

Gaussian Process Regression Gaussian Process Regression (GP) models are based on a biased multi-variate normal distribution, where the prediction is the expected value according to the maximum likelihood principle. They are beneficial since they are one of the few regression algorithms with a probabilistic prediction, therefore giving a confidence of the prediction. The used kernel is ($0.5 \times$ rational quadratic kernel + white kernel). This kernel combination is the result of a grid search including multiple kernels implemented in [33].

Decision Trees with Adaptive Boosting Decision Trees (DT) are a common interpretable model, which require little data preprocessing. In our application, we use DTs with the ensemble method Adaptive Boosting [35], which has been shown to improve DT regression performance. As a result of a hyperparameter optimization run the minimum number of samples per leaf of the DT base model is set to 5, the number of estimators for the Adaptive Boosting is set to 500 and the learning rate to 1×10^{-7}.

2.4. Neural networks

The Neural Network (NN) model has the capability of highly nonlinear modeling. We created the NN model with PyTorch [36] version 1.7.1. Neural Networks are applied in various disciplines and have already shown good results in soiling determination for PV [5–7].

We chose a feed-forward architecture with an input layer of 10 to 29 inputs according to the dataset, with five hidden layers containing 512, 256, 128, 64, and 32 neurons (see Fig. 3). The number of hidden layers and the number of neurons in the first hidden layer were determined via an architecture optimization in a preliminary study (see Appendix B). The number of neurons in the subsequent layers are steadily decreased to match the output layer. Since we have a regression problem, we have a single output neuron for the soiling prediction. We include dropout with a constant probability of 0.4 applied to all layers to prevent overfitting. To speed up model convergence, we initialized the output layer with the average gloss value of the training dataset. In order to optimize the hyper-parameters of the Neural Network, a grid search with three activation functions (Rectified Linear Unit (ReLU), Sigmoid, Hyperbolic tangent (Tanh)) and three learning rates (3×10^{-4}, 6×10^{-4}, 9×10^{-4}) was performed. Early

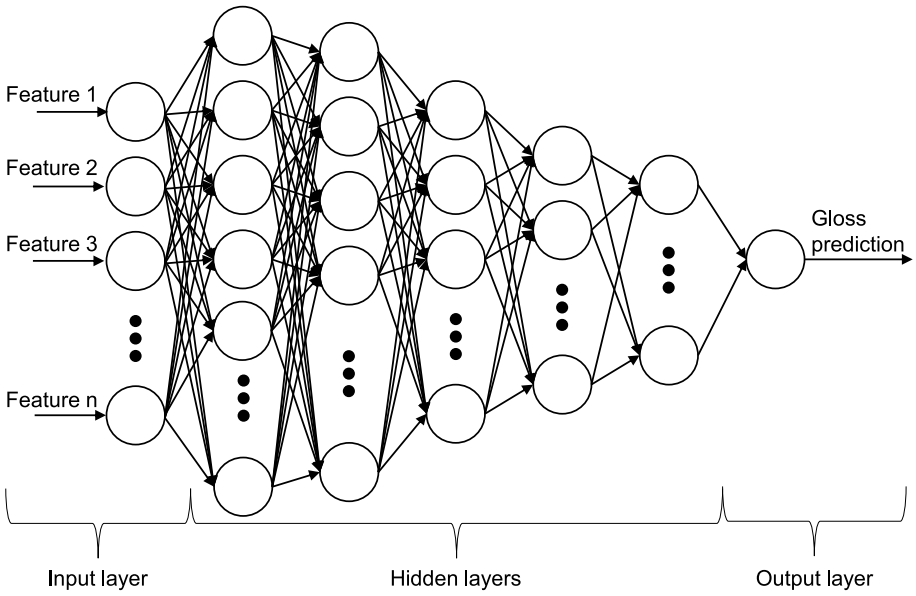

Fig. 3. Schematic of Neural Network architecture with number of input features n according to dataset (n=10, n=11, n=14, n=29), five hidden layers, dropout applied to all layers, and gloss prediction as model output.

stopping was applied if the validation loss of the current epoch was higher than the last 1000 validation losses and at the same time the number of the current epoch is 1000 higher than the epoch with the best results so far.

3. Metrics

Metrics for model comparison are used to decide which of the presented models shows the best results with one of the created datasets. According to the use case, we defined two different types of metrics. The first type focuses on the regression problem, where the gloss measurements serve as direct comparison to the model predictions. In the best case the gloss measurements can be directly replaced by the predictions from the model.

From an application point of view, the model predictions can provide information on the need for cleaning for individual collectors in the field. This second type of metric is used for the classification problem, where we simply want to decide whether the mirror soiling is above or below a certain gloss threshold. In that case we have an easier classification problem with the potential to further optimize the cleaning schedule.

The models are the same for both evaluation metrics. The division into two metrics should give different insights into the model prediction quality.

3.1. Prediction of gloss values (regression problem)

We use four different metrics to decide which models are best suited for the regression problem of replacing gloss measurements. The coefficient of determination (R^2) is a commonly used metric to quantify how well a model fits the underlying measured values. If the model perfectly fits the measured values R^2 equals 1. If there is no relation between model and measurement R^2 equals 0. Mean squared error (MSE) is especially useful to give outliers a higher weight since it takes the square of the difference between measured and predicted value. Root mean squared error (RMSE) takes the square root of MSE and therefore has the advantage of conserving the unit of the value. Mean absolute error (MAE) takes the average of all differences between measured and predicted value without squaring it. Outliers are not penalized stronger than other values.

3.2. Prediction of cleaning necessity (classification problem)

In order to evaluate the quality of the classification task we define a cleaning threshold that is set at 90 gloss units. This value is also used for cleanliness thresholds in Wolfertstetter et al. [37] and Pettit et al. [38]. Collectors with higher gloss values do not need to be cleaned, collectors with gloss values below this limit need to be cleaned. The model predictions and gloss measurements are compared with the cleaning threshold in a confusion matrix including true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), see Table 2.

To compare these results with the current cleaning procedure at Andasol-3 a similar confusion matrix is designed (Table 3). The current cleaning procedure is based on a fixed-time schedule where a collector is cleaned approximately every four days. The recommendation to clean is taken by applying the cleaning threshold to the measured gloss values, and compared with the current cleaning procedure.

Additional calculated metrics from the confusion matrix are accuracy, recall, precision, and F1-score, shown in Eqs. (2)–(5). Accuracy describes the ability to predict correctly if the cleaning is useful or not and therefore should be as high as possible. Recall describes the ability to find useful cleanings and should be as high as possible. Precision again should be maximized and can be described by the ability not to clean too early. F1-score combines precision and recall with the harmonic mean. This has the advantage of not being sensitive to class imbalances, as accuracy is. Therefore, F1-score is assumed to be the most important metric for this work since we have an imbalanced dataset (see Fig. B.20 in the appendix). The other metrics are nonetheless given for completeness.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Table 2
Confusion matrix definition for evaluation of gloss predictions (here referred to as prediction). The models recommend or do not recommend a collector cleaning. This is compared to the actual necessity given by the measured gloss (here referred to as measurement) above or below the cleaning threshold (here referred to as threshold).

True positives (TP) (prediction < threshold) & (measurement < threshold) → cleaning recommended → cleaning necessary	**False positives (FP)** (prediction < threshold) & (measurement > threshold) → cleaning recommended → cleaning unnecessary
False negatives (FN) (prediction > threshold) & (measurement < threshold) → cleaning not recommended → cleaning necessary	**True negatives (TN)** (prediction > threshold) & (measurement > threshold) → cleaning not recommended → cleaning unnecessary

Table 3
Confusion matrix definition for comparison of gloss measurements (here referred to as measurement) and actual cleaning activities. Cleaning is required with gloss measured below the cleaning threshold (here referred to as threshold). Actual cleaning procedure is determined from collector positioning commands.

True positives (TP) (measurements < threshold) & collector is cleaned → cleaning recommended → collector cleaned	**False positives (FP)** (measurements > threshold) & collector is cleaned → cleaning not recommended → collector cleaned
False negatives (FN) (measurements < threshold) & collector is not cleaned → cleaning recommended → collector not cleaned	**True negatives (TN)** (measurements > threshold) & collector is not cleaned → cleaning not recommended → collector not cleaned

3.3. Permutation feature importance

To gain insight into how useful each model input is for model predictions, we calculate a permutation feature importance [39]. It is calculated by randomly shuffling a single feature between samples and reevaluating the MSE of the model with the shuffled feature. This is performed repeatedly for every feature individually. With this procedure the relationship between feature and target value is removed. The increase in the model loss then corresponds to the importance of the shuffled feature. One major drawback is that permutation feature importance is insufficient for correlated features. This can result in the calculation of a too low loss and erroneously lead to the assessment of low importance for the correlated features. As a consequence, high feature permutation losses stand for a high feature importance, but low feature permutation losses cannot be interpreted as low feature importance.

3.4. Model uncertainty

In order to ensure the significance of the results, we estimate the model uncertainty by splitting the uncertainty calculation into two parts, systematic uncertainty and statistical uncertainty. The systematic uncertainty, when applied to the classification task, showed very small values of less than 0.1 %. We therefore assume they can be neglected and only statistical uncertainties are calculated for the classification task. For the regression task we only calculated the systematic uncertainty. The statistical uncertainty is only applicable if we have results from a confusion matrix to calculate the uncertainties.

3.4.1. Systematic uncertainty

A major source of systematic uncertainty comes from the particular random initialization of a model used. This is applicable for the Neural Network and Decision Tree models. In order to estimate the influence of the random initialization, 30 runs for each model with different initializations were performed. Every regression metric for Neural Network and Decision Tree is an average value supplemented with the uncertainty $u(x)$ from Eq. (6).

For the Gaussian Processes, Support Vector Regression, and Linear Regression model no systematic uncertainty is calculated.

$$u(x) = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \tag{6}$$

3.4.2. Statistical uncertainty

The statistical uncertainty arises due to the fact that the datasets contain only a limited number samples, which therefore cannot perfectly describe the underlying data distribution. We use the Clopper–Pearson (CP) interval [40] with a coverage rate of 95 % to calculate the binomial confidence intervals. The CP interval is a conservative method, since it is always above the nominal coverage.

4. Results and discussion

In the first part of this section, the results for the presented models with different datasets are discussed and compared. The performance of all models was investigated for two types of metrics. Regression metrics were used to investigate the ability to replace gloss measurements, and classification metrics evaluated the ability to accurately detect whether a collector should be cleaned or not. In the second part of the results the most accurate model from the regression task is used and compared to subfield average gloss values. In the last part of the results the feature importance is discussed throughout all used models.

4.1. Model preselection

In order to find the best suited model for each of the datasets, all models are trained with every prepared dataset. The metrics from Section 3 are used to compare the model performance on the test data. The test dataset was only used for the purpose of model comparison and was not involved in model training or model optimization. Besides finding the best model for every dataset, we investigate if classical machine learning models from Section 2.3 are sufficient to deliver accurate results. The use of datasets with different numbers of included features gives information whether a small measurement setup at a power plant with a lower number of data sources is sufficient to deliver enough information for the purpose of soiling determination. Results from the regression metrics throughout different models and datasets are shown in Fig. 4. For the Neural Network model, the best results from the grid search are included. Values in parentheses are the systematic uncertainty values for Neural Network and Decision Tree models.

4.1.1. Prediction of gloss values (regression problem)

The Linear Regression model shows the highest $R^2 = 0.34$ and lowest $MSE = 17.46$, $RMSE = 4.18$, and $MAE = 3.01$ with `dataset_3`. Linear Regression shows the worst results in the regression task for all compared models.

The Support Vector Regression model shows the best results with `dataset_3` for the metrics $R^2 = 0.52$, $MSE = 12.76$, $RMSE = 3.57$ and $MAE = 2.21$. However, these high losses are considerably worse compared to the NN, DT, and GP.

In the Gaussian Process regression model, the best results are achieved with `dataset_3`, with $R^2 = 0.73$, $MSE = 7.08$, $RMSE = 2.66$, and $MAE = 1.68$. The results show an improvement if bigger datasets are used.

The results of the grid search for the Neural Network model are shown in Fig. B.21 in the appendix. The figure shows the averaged calculated regression metrics for different random seeds together with the systematic uncertainty ($u(x)$) for each metric. The maximum $R^2 = 0.74$ and minimum $MSE = 6.93$ and $RMSE = 2.63$ are achieved with the smallest dataset, `dataset_0`, Tanh activation and a learning rate

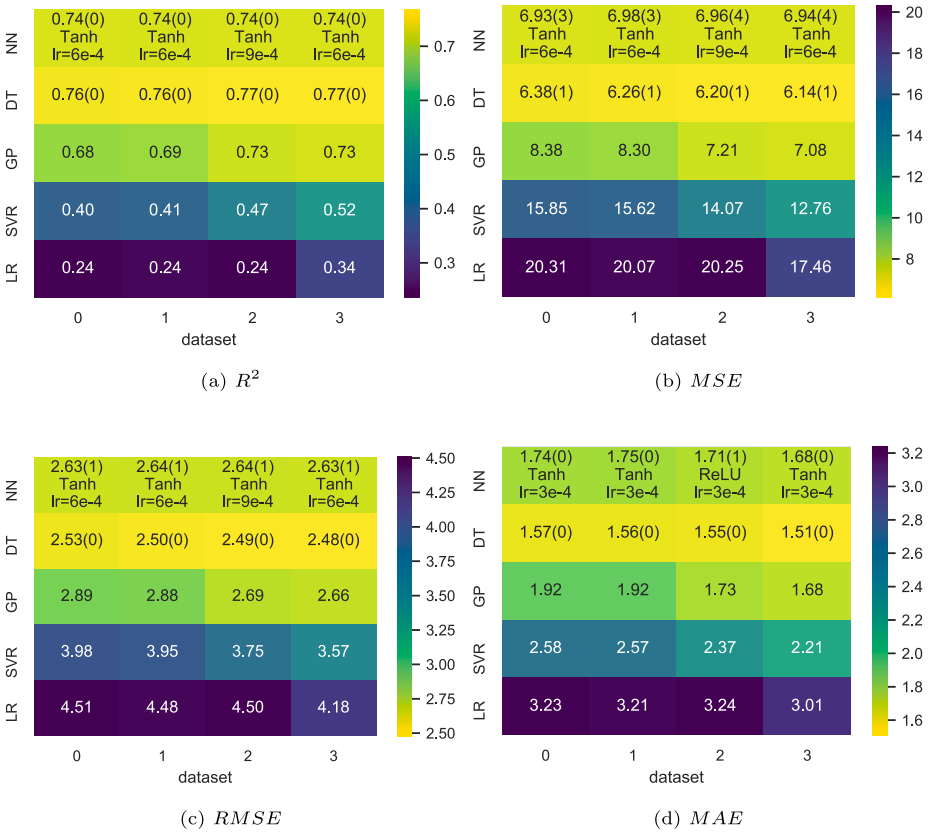

Fig. 4. Model and dataset comparison for regression problem with R^2, MSE, $RMSE$, and MAE metrics. The Decision Tree (DT) and Neural Network (NN) show the best results, followed by the Gaussian Processes (GP) model.

of 6×10^{-4}. However, the results from `dataset_3` are comparable within the range of uncertainty. The minimum $MAE = 1.68$ is achieved with the largest dataset, `dataset_3`, Tanh activation and a learning rate of 3×10^{-4}. As an overall trend, it can be seen that Tanh activation performs best for each dataset, consistently outperforming ReLU and Sigmoid activation. In addition, the results for R^2, MSE, and $RMSE$ with Tanh activation are similar throughout different datasets, only MAE is reduced with a bigger dataset, e.g. `dataset_3`. The regression results for the best configuration from the Neural Network model grid search are shown in the first rows of Fig. 4. The model achieves the second best results compared to the other models with a near consistent performance for the different datasets.

The best regression results for all investigated models are achieved with the Decision Tree model with a maximum $R^2 = 0.77$, minimum $MSE = 6.14$, $RMSE = 2.48$, and $MAE = 1.51$ with the largest dataset, `dataset_3`, see Fig. 4. A slight improvement in model performance can be seen with the use of a bigger dataset, from `dataset_0` to `dataset_3`.

MSE of train, validation, and test dataset part are compared in Table 4. All models are included with their configuration that achieves the best results. MSE of validation is lower that the MSE of test data. This indicates that all models are slightly overfitted.

To sum up, the best results are achieved with the Decision Tree model with $R^2 = 0.77$. This coefficient of determination can be interpreted as percentage value, where 77 % of the data can be explained with that model. Second and third best results are reached with Neural Network and Gaussian Process. Support Vector Regression and Linear Regression models show clearly worse results.

Scatter plots of the predicted and measured gloss values are shown in Fig. 5. Predictions from Support Vector Regression and Linear Regression models show high deviations from the measurements especially in the area with measured gloss values below 90. Decision Tree and Neural Network models have a nearly constant performance for different datasets, which can be interpreted as a high robustness against a changing measurement data availability. These models seem to be most suitable for the use in different power plants with a varying measurement setup.

Table 4
MSE of train, validation, and test dataset part. All models show small difference of MSE of validation and test data. This shows that the models are slightly overfitted.

Model	Train	Validation	Test	Model parameter
NN	4.59	6.43	6.93	dataset_0, Tanh, lr=6e−4
DT	1.66	5.72	6.14	dataset_3
GP	2.53	6.35	7.08	dataset_3
SVR	13.65	11.98	12.76	dataset_3
LR	18.87	16.85	17.46	dataset_3

4.1.2. Prediction of cleaning necessity (classification problem)

Linear Regression achieves the highest true negatives and lowest false positives with `dataset_0`. The highest Accuracy and Precision with `dataset_1`. Highest true positives, Recall, and F1-score, and lowest false negatives with `dataset_3`. The classification results for Linear Regression are the worst within the comparison, but compared to the current cleaning schedule the model still shows better results.

The best classification results for the Support Vector Regression model are obtained with `dataset_2` for true negatives and false positives. All other classification metrics show their optimum with `dataset_3`. Precision values stand out from the other metrics. Here the Support Vector Regression model has the highest score. This can mainly be explained by the low false positive rate and the calculation of Precision, see Fig. 6 and Eq. (4). The high Precision value shows a good ability not to clean too early. This might be beneficial when cleaning is very costly and a high solar yield is not important.

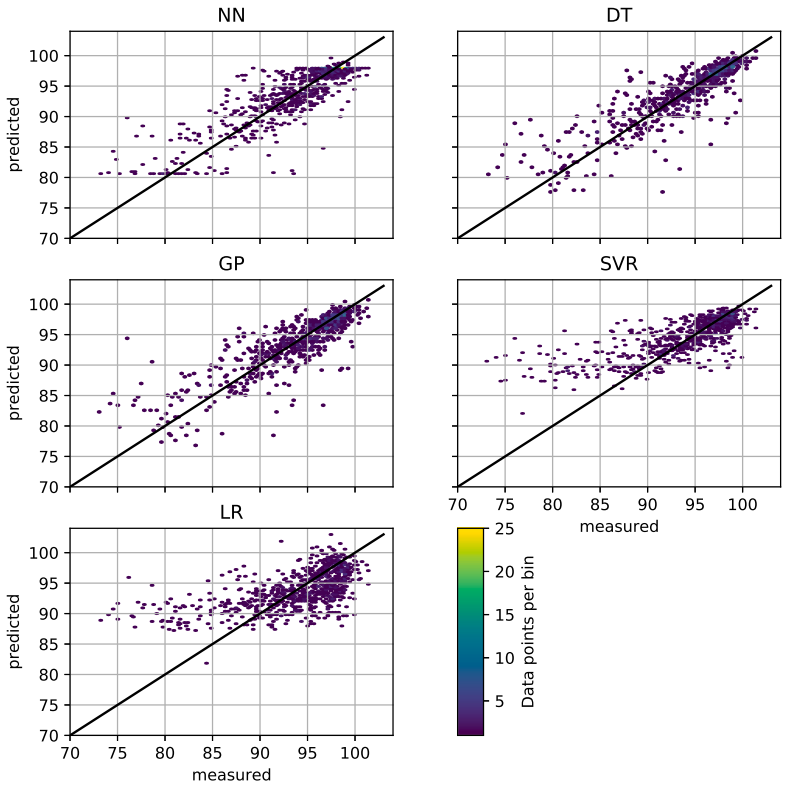

Fig. 5. Scatter plot of gloss measurements and predicted values for all models and datasets which showed the highest R^2 value. Support Vector Regression and Linear Regression model show high deviations for measured gloss values below 90 and are therefore not applicable to solve the regression problem. Colors of data points indicate the number of data in this bin.

The classification results for Gaussian Processes show a heterogeneous picture regarding the different datasets. The highest true negative, highest precision and lowest false positives as well as the highest Accuracy and Precision are obtained with `dataset_1`. Highest true positives, highest recall and f1-score, and lowest false negatives are achieved with `dataset_2`.

The results of the grid search for the Neural Network model for classification are shown in Figs. B.22 and B.23 in the appendix. Accuracy, Recall, Precision and F1-score show small changes with varying learning rates, datasets and activation functions. In most cases, Tanh activation shows the best results throughout different datasets. Comparable results can be seen for true positives, false positives, false negatives, and true negatives in Fig. B.22 in the appendix. The classification results for the best configuration from the Neural Network model grid search compared to the other models is shown in Figs. 6 and 7. The Neural Network model achieves one of the best results compared to the other models with a nearly consistent performance for different datasets. The adaptability of the Neural Net to different datasets can be seen as a indicator of the advantage of using it in other power plants, with a possibly different measurement data availability.

The Decision Tree model show the highest true negative, true positive, Accuracy, Recall, Precision, and F1-score, as well as the lowest false positive and false negative rate for `dataset_3` (see Fig. 6). With the exception of precision these are the best values compared to the other models. The Decision Tree is slightly more sensitive to the usage of different datasets compared to the Neural Network. The Decision Tree is therefore the most reliable model in predicting correctly whether the collector should be cleaned or whether the cleaning is not yet necessary. This balance seems to be advantageous for most parabolic trough power plants, which try to find a good compromise between low cleaning effort and high solar yield.

In conclusion, from the classification metrics we can see that the Decision Tree, Neural Network, and Gaussian Process model have similar results. The three models show closer results than in the regression task. This makes sense as the classification task does not require a precise gloss prediction but only a classification to be made by the models. The classification task thus allows for a margin of error around the threshold. Support Vector Regression and Linear Regression models show worse results, but even these poor results of the models are advantageous over the current cleaning schedule.

Since the Decision Tree shows the best results, its predictions are compared with the current fixed-time cleaning schedule. The definitions for the confusion matrix evaluation were given in Table 3. Similar metrics are calculated and the results collected in Fig. 8. The poor results from the current cleaning schedule indicate that the available gloss measurements do not seem to be used frequently for the decision about whether the collectors need to be cleaned or not. As stated in Section 3.2, we can compare these results with the classification results from the model predictions. Relevant differences can be seen in the false positives rate. This can be interpreted as the model, when compared to the current schedule, not proposing a cleaning procedure too early. Therefore, unnecessary cleanings can be reduced by 14.3 %. The clearly higher true positives rate indicates a much better ability to detect situations where a cleaning procedure is necessary. With the use of the model predictions, necessary cleanings are detected additionally in 12.2 % of the investigated cases. The increase of Accuracy can be interpreted as an overall increased cleaning prediction quality. These results clearly show the advances of a condition based cleaning strategy, compared to a fixed cleaning frequency. This is also in accordance with other studies, where a condition based cleaning strategy is compared to a fixed-time strategy [41–43].

4.2. Comparing subfield average gloss with model predictions

A common practice for monitoring the mirror cleanliness is a sampling method which delivers an average cleanliness value for the subfield or entire plant. An example for such a procedure with readings

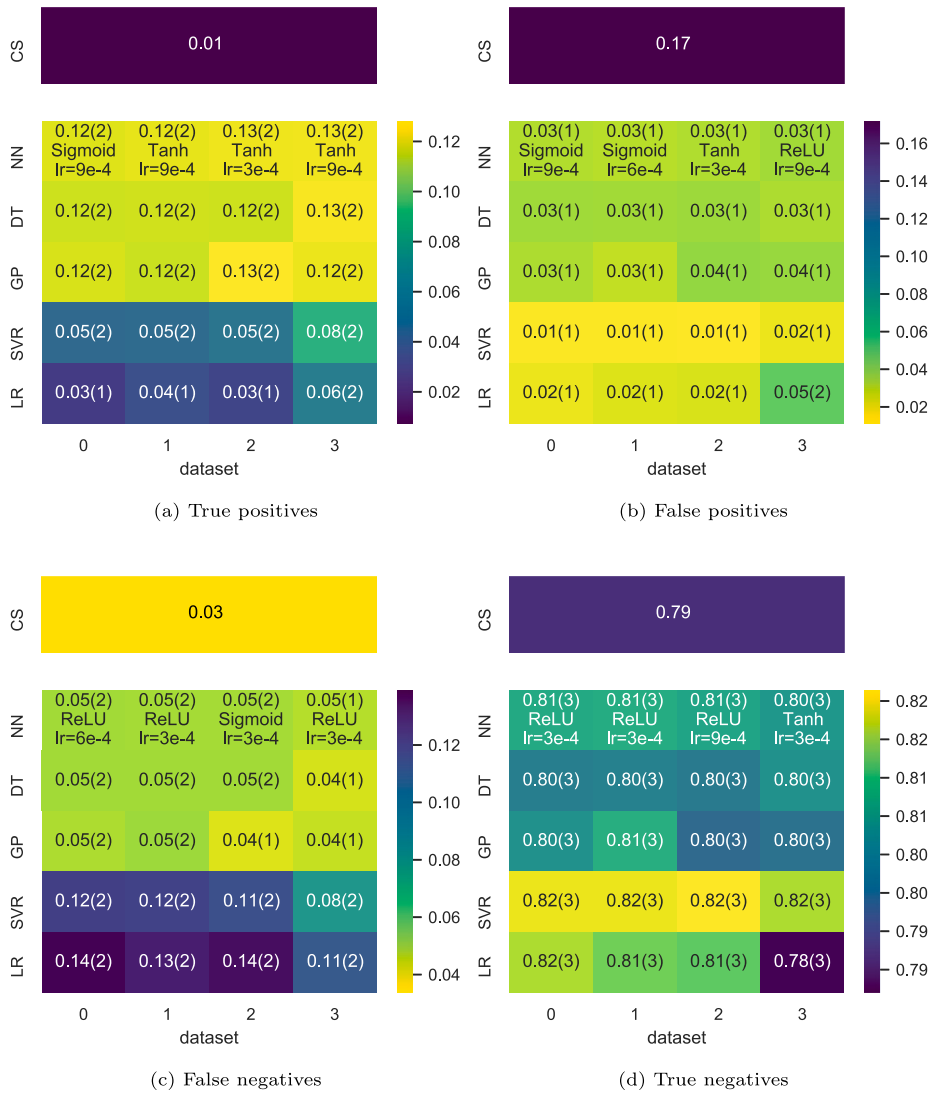

Fig. 6. Model and dataset comparison for confusion matrix results true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN). Color scales vary between metrics for clarity, in all cases lighter is better. The metrics for the current cleaning schedule (CS) are included for comparison. Neural Net, Decision Tree, and Gaussian Processes show better results for cases where cleaning is necessary (TP, FN), Support Vector Regression and Linear Regression are beneficial for cases where cleaning is not necessary (FP, TN).

taken at 84 locations is given in [27]. For the comparison, the model which showed the best regression results is used (Decision Tree trained with `dataset_3`). Data from the test subset are used for the comparison. We require a minimum of two gloss measurements within an entire subfield to be used for predictions. A total of 188 different subfield average values are included in the comparison subset. We observe at most only ten gloss measurements for an entire subfield. With this low number of samples a student t distribution with a 68.3 % confidence interval is used to estimate the uncertainty of the average values.

Results are shown in Fig. 9. We see that in most cases the model fits the average gloss measurements. However, we see a high measurement uncertainty for both the gloss measurements and gloss predictions. This is mainly caused by the low number of samples per subfield and a high variation within the subfield. The latter is especially the case for gloss measurements.

One of the main reasons for the high uncertainties is caused by the usage of the test dataset. It was randomly sampled from the dataset and it is therefore very unlikely that an entire subfield measurement campaign is contained in the test dataset. To gain more meaningful information about the capabilities of the model to predict average values, a new dataset with coherent measurements is needed. Moreover, in the current model information of neighboring collectors is not used for the individual soiling determination. Predicting average values would benefit from a model, which processes data from the entire subfield collectively.

4.3. Permutation feature importance

In order to get insights into the usefulness of every model input for the model prediction, we calculated the permutation feature importance for every input and model. Fig. 10 shows the feature permutation losses minus the MSE of the original model on the x-axis for every feature on the y-axis. The results show high losses, and therefore high feature importance, for generally the same features across all models. These are sine and cosine of week of year, irradiance features, wind features, precipitation maximum, and hours since precipitation cleaning threshold was reached. Linear Regression, Neural Net, and Gaussian Processes reach higher importance for the features collector temperature and wind u, v average. Particle concentration feature (PM_{10}) was especially useful for Neural Net, Gaussian Processes, and Support Vector Regression model.

In conclusion, permutation feature importance shows a high seasonal influence for soiling for all models, and for the majority of the

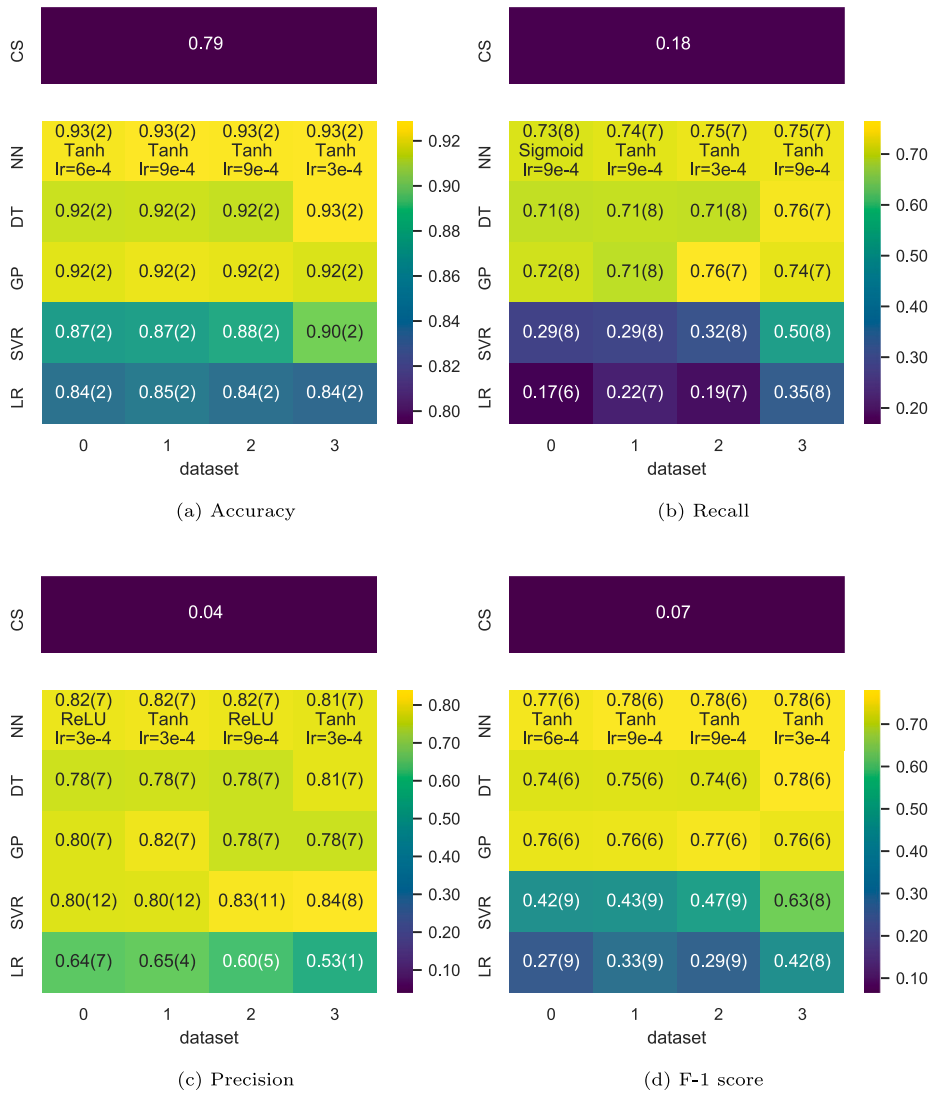

Fig. 7. Model and dataset comparison for confusion matrix results Accuracy, Recall, Precision and F1-score. Neural Net, Decision Tree, and Gaussian Processes show the overall best results for correctly predicting if the collector should be cleaned or not. The metrics for the current cleaning schedule (CS) are included for comparison.

models a high influence from irradiance, wind, and PM_{10}. The high seasonal influence is also reported in [27], PM_{10} showed the highest correlation with soiling of PV modules in [29]. However, the characteristic of the used method makes it difficult to determine without doubt which feature shows a low importance.

5. Conclusion and outlook

In this work, a method for soiling determination using the operational data from the solar field and additional meteorological data was presented. The method delivers helpful information on the current soiling status, which can be used for optimization of the cleaning schedule of the parabolic trough field. Using one of the explored models for soiling determination, therefore, helps to apply an economical, optimized cleaning strategy in parabolic trough fields. One drawback of these investigations is the use of gloss measurements and their application as soiling indicator. Although it is used in some power plants as a quick and simple measurement technique for spatially resolved soiling evaluations, the use of reflectance measurements would be preferable.

A first evaluation criterion for the model results is the direct comparison with gloss measurements. In this regression problem, the best results are again achieved with the Decision Tree model and the usage of the largest dataset. 77 % of the test data can be explained with the predictions of this model. Slightly worse results are achieved with Neural Network and Gaussian Processes models. This result is in accordance with Grinsztajn et al. [44], who discusses the superiority of tree-based models over Neural Networks.

The key evaluation criterion for the model results is the prediction of cleaning necessity and the comparison to the current fixed-time cleaning schedule. All investigated models outperformed the current fixed-time cleaning schedule. As a main result, we showed a potential reduction of unnecessary cleanings by 14.3 % and increased detected necessary cleanings by 12.2 % in investigated cases. This is achieved by using the Decision Tree model with the largest training dataset. The results from Neural Network and Gaussian Processes models show similar performance. A detailed comparison of the economic benefits of a cleaning schedule based on soiling thresholds and a fixed-time cleaning schedule is shown by Wolfertstetter et al. [37]. They showed a relative profit increase of the threshold-based cleaning schedule over the fixed-time schedule of 0.36 %.

The comparison of subfield average values to Decision Tree predictions shows in most cases a good fit between measurements and predictions but with high uncertainties. Testing the model with an specially created test dataset with coherent measurement within one subfield would be needed to lower the uncertainties.

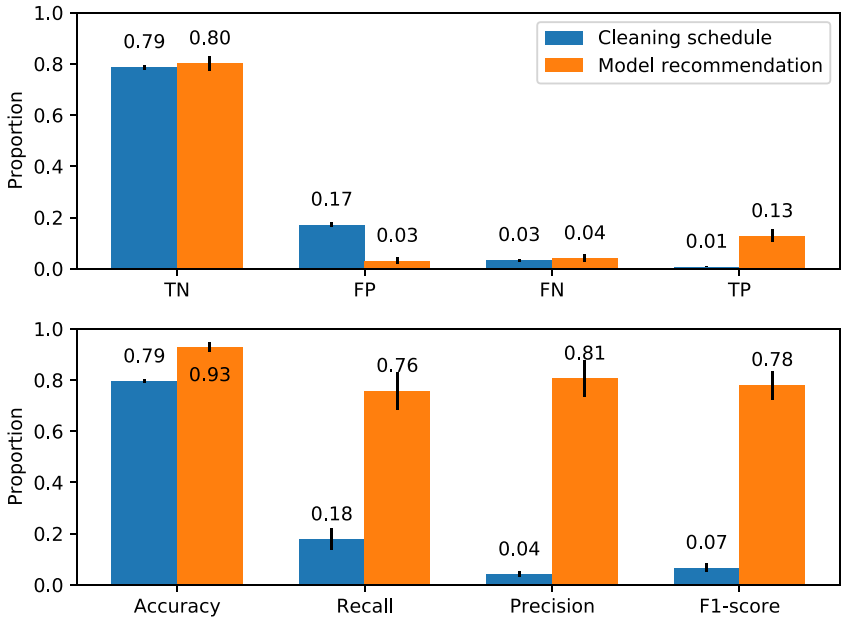

Fig. 8. Comparison of current cleaning schedule with model recommendations from Decision Tree and `dataset_3`. The Decision Tree shows an improvement in almost all cases compared to the fixed-time cleaning schedule. Unnecessary cleanings can be reduced by 14.3 % (FP), necessary cleanings are detected additionally in 12.2 % (TP).

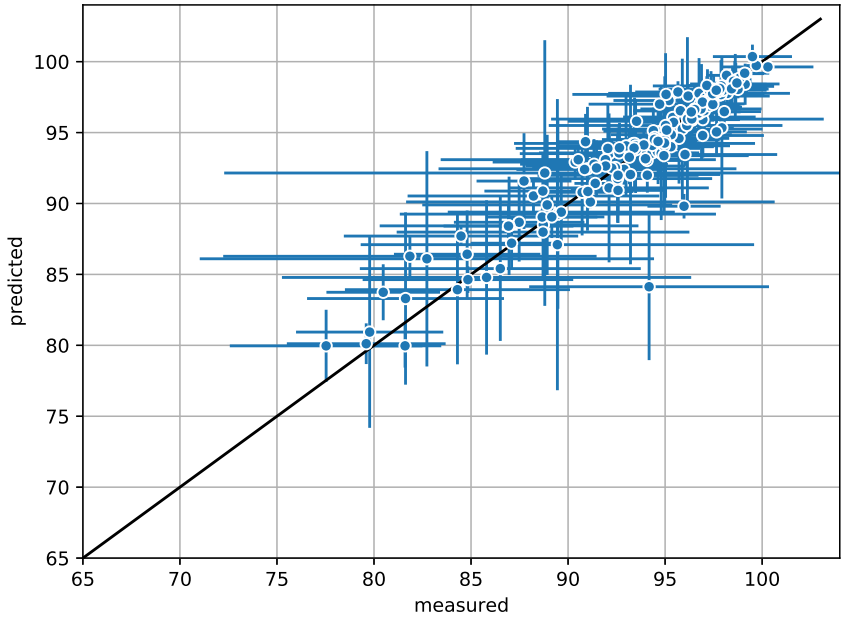

Fig. 9. Comparison of average subfield values from gloss measurements (horizontal error bars) and Decision Tree model predictions (vertical error bars). The model fits well the actual average subfield gloss especially at high gloss values.

All used models showed a high seasonal influence on the soiling prediction. Other important inputs for the majority of the models are irradiance, wind and particle concentration (PM_{10}). Nevertheless, using the permutation feature importance method makes it difficult to concretely determine which features do not have a high importance.

The low variations in regression results between different datasets of the Decision Tree and Neural Network model makes them most suitable for the application in a different power plant with a varying measurement setup. Following the success of transfer learning in other domains [45], we believe that the fine-tuning of the trained models to a new power plant requires much less data than the initial training and therefore only a small amount of additional gloss measurements have to be taken at the new power plant for our models to provide benefit. Due to the limited amount of data we made the design choice of using only 10% of the data for validation and test dataset. For future investigations using cross-validation would be a promising approach to better estimate the training variance of the relatively small dataset we used. By making the dataset publicly available, others can combine it with their own data to create a more comprehensive training, validation, and test set.

Further future developments with a different scope may focus on the determination of whole subfield average soiling values. Convolution [46] or attention-based neural networks [23], for example, may be an option for this purpose as they can consider information from multiple collectors rather than predicting a soiling for each collector independently.

Selecting meaningful time-series features was a manual task in this work. This goes along with a simplification via the compressing of

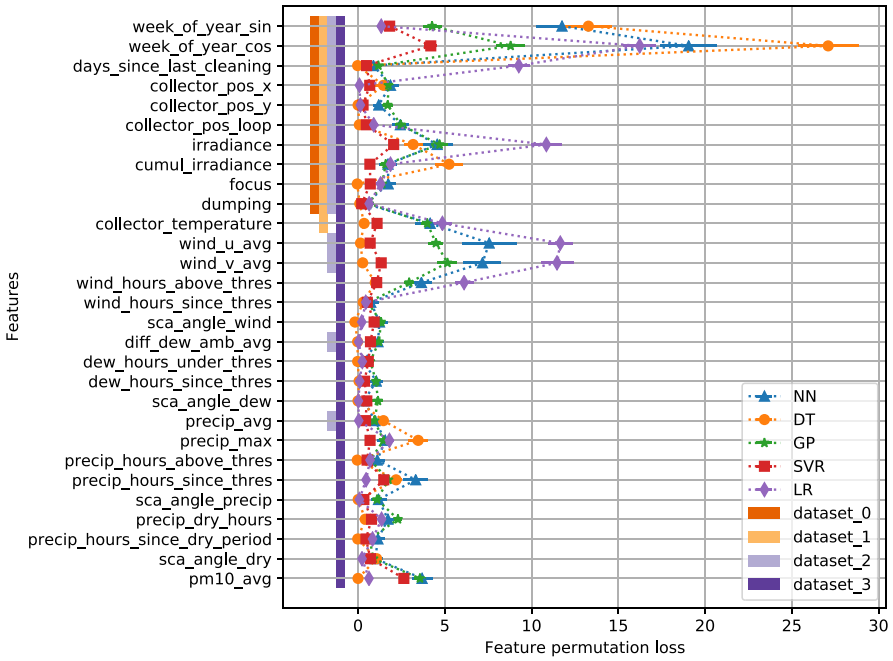

Fig. 10. Permutation feature importance for all models across different features. A high loss indicates a high feature importance. All models show a high seasonal influence for soiling estimation. The majority of the models also show a high influence from irradiance, wind, and PM_{10} features. The colored bars on the left of the figure indicate to which dataset the features belong to.

time-series data into a fixed number of values with the usage of mean, minimum, maximum, and time duration values. Using transformer neural networks [23], for example, can be a way to automate the engineering of relevant features and avoid compressing the data more than necessary. Their built-in attention mechanism may be a promising way of extracting meaningful features, which help the model find the best prediction.

The datasets are created with the assumption of no concept drift, i.e. that the properties of the dataset do not change over time. Namely, we assume that the factors determining soiling are consistent from year to year. If this assumption holds, a model trained on historical data can be expected to perform on future data. We leave the investigation of concept drift as an avenue for future work.

In this work, we have shown that machine learning models, when utilizing readily available operational and meteorological data, are capable of delivering valuable information about the soiling levels within a solar field. This approach, when applied to parabolic trough solar fields, has the potential to reduce operational and maintenance costs, increase energy yields, and extend the lifespan of the field as a whole. Determining soiling using this approach can also be used for future anomaly detection methods. Soiling overlays all other performance degrading effects and it occurs in every solar field. It is therefore necessary to know the current soiling level to detect other anomalies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the Helmholtz Association Initiative and Networking Fund, Germany under the Helmholtz AI platform grant. The authors want to thank Marquesado Solar for providing raw data from the Andasol-3 power plant and for their support during the development of the model.

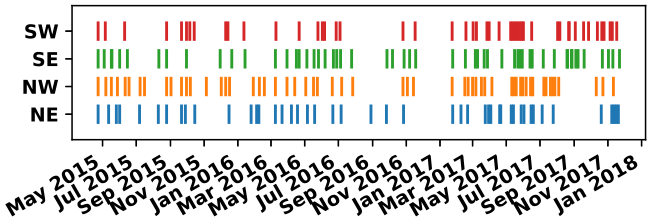

Fig. A.11. Timestamps of gloss measurements for each subfield (SW: southwest, SE: southeast, NW: northwest, NE: northeast) from 2015 to 2018. In most cases, the data used show a uniform distribution throughout the year, except for the period between December and January, for which data are only available in 2015.

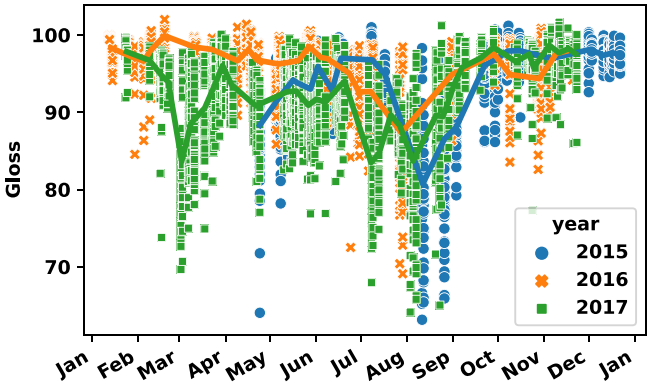

Fig. A.12. Annual pattern of gloss values used in the data sets from 2015 to 2017. The dots show the measurements included in the data set. The lines are weekly average values. Seasonal influences that reduce gloss values can be seen in the summer months of July through September.

Appendix A. Representation of input data

See Figs. A.11–A.17.

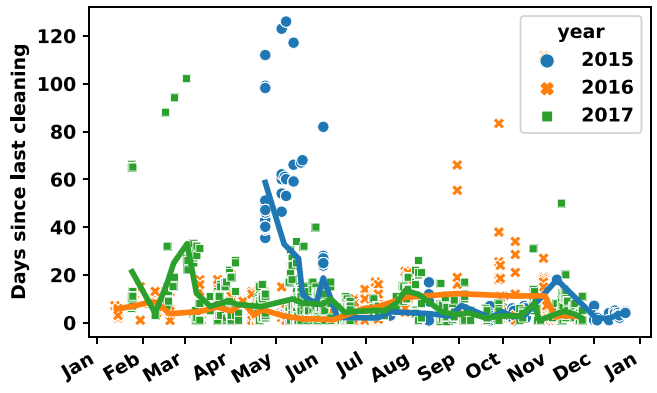

Fig. A.13. Annual pattern of days since the last cleaning of the collectors used in the datasets from 2015 to 2017. The dots show the measurements included in the dataset. The lines are weekly averages. Median values for each year range from 3 to 6 days.

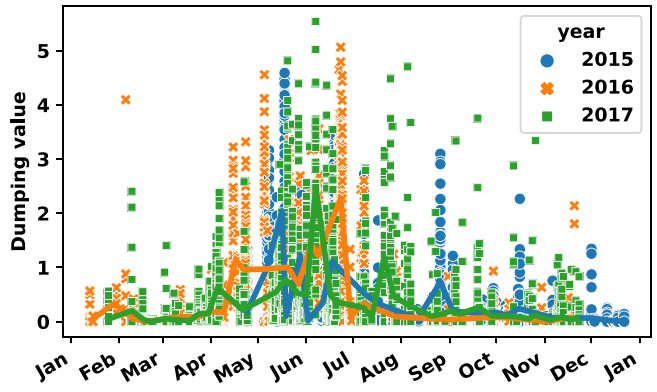

Fig. A.14. Annual pattern of dumping values used in the datasets from 2015 to 2017. The dots show the measurements included in the dataset. The lines are weekly average values. Due to high solar irradiation, the solar field has higher dumping values in the summer months.

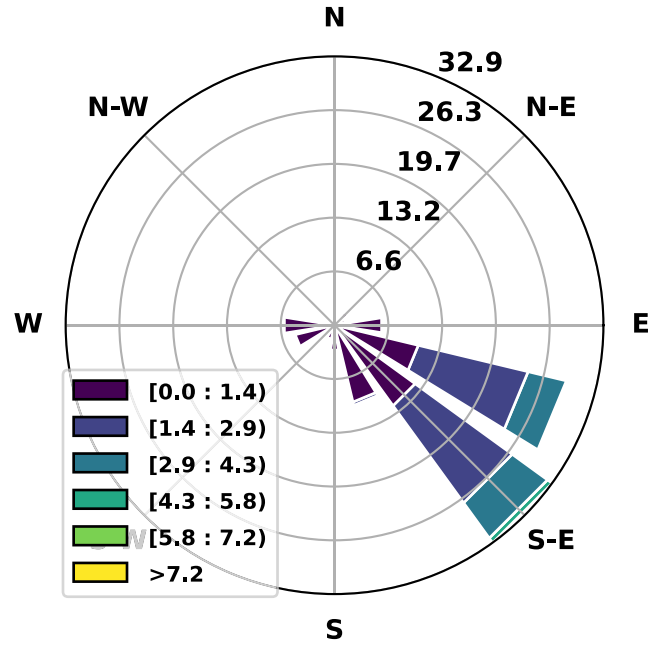

Fig. A.15. Wind rose for the data used in this paper. The main wind direction is southeast, according to the terrain characteristics.

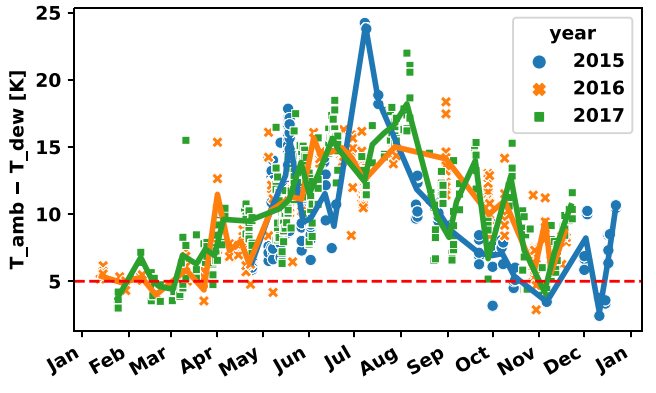

Fig. A.16. Difference between the dew point temperature and the ambient temperature. The dots show the measurements included in the data set. The lines are weekly averages. The horizontal dashed red line shows the theoretical threshold used for dew formation. Dew formation occurs mainly between October and April. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

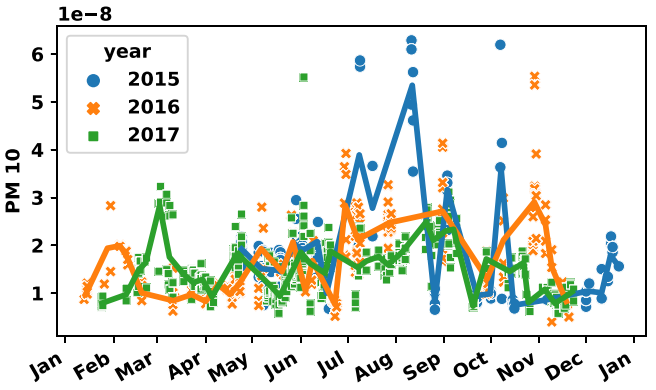

Fig. A.17. Annual pattern of PM_{10} values used in the datasets from 2015 to 2017. The dots show the measurements included in the dataset. The lines are weekly average values. The highest PM_{10} values can be seen in the summer months, which is consistent with the lower gloss values in those months.

Table B.5
First part of architecture optimization with fixed number of neurons in first layer (1024) and varying number of hidden layers. Lowest MSE is achieved with 5 hidden layers.

Number of hidden layers	1	2	3	4	5	6
MSE	8.73	6.66	6.05	6.16	**6.04**	6.1

Table B.6
Second part of architecture optimization with fixed number of layers (5) and varying number of neurons in first hidden layer. Lowest MSE is achieved with 1024 neurons in first hidden layer, but with highest relative runtime.

Number of neurons in first hidden layer	64	128	256	512	1024
MSE	12.93	9.93	8.40	6.75	**6.04**
relative runtime	**0.51**	0.57	0.65	0.70	1.00

Appendix B. Neural network architecture optimization

The number of hidden layers and the number of neurons in the first hidden layer were determined with a two stage optimization procedure. It was performed with `dataset_0`, a learning rate of 9×10^{-4}, and Tanh activation trained for maximum 3000 epochs with early stopping enabled. In the first stage, the number of hidden layers was determined, with the number of neurons in the first hidden layer fixed to 1024. The lowest MSE was achieved with 5 hidden layers (see Table B.5). Fixing the number of hidden layers to 5, the number of neurons in the first hidden layer was determined in the second stage of the optimization

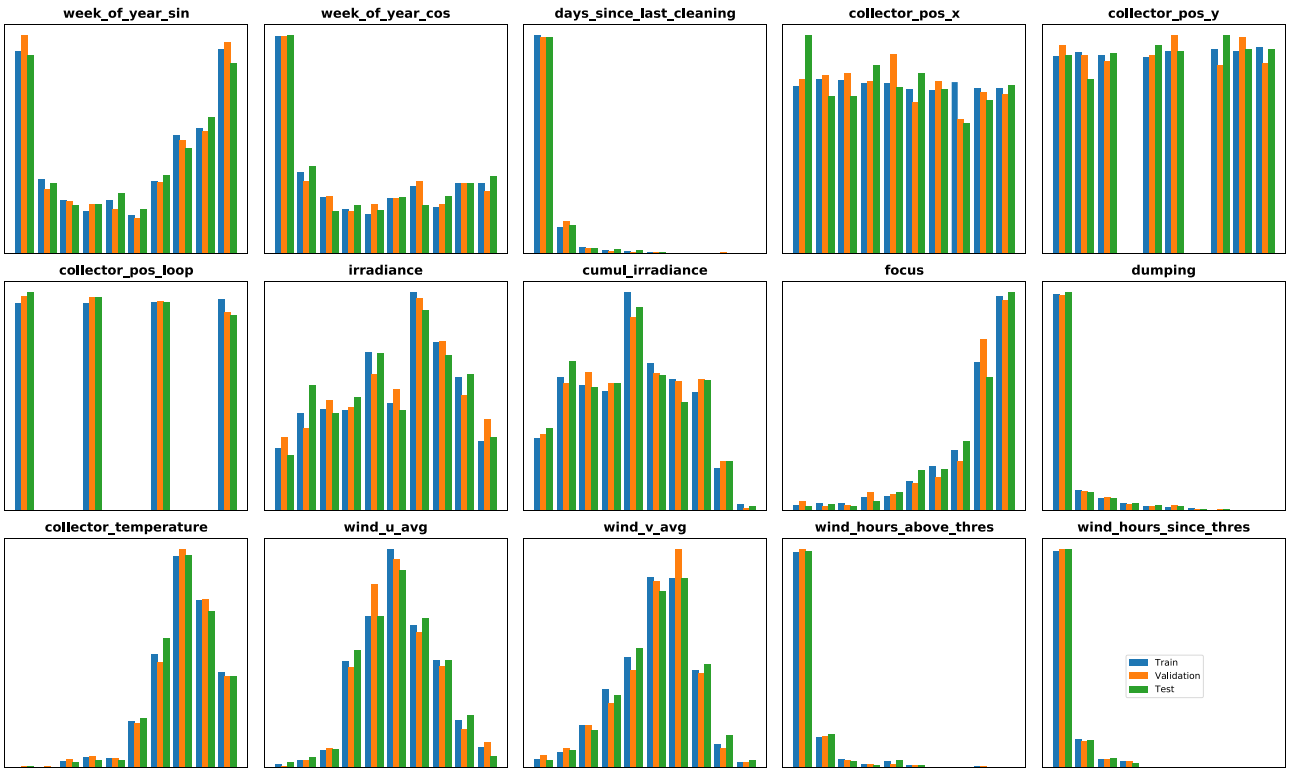

Fig. B.18. Histogram plots (part 1) of training (blue), validation (orange) and test (green) distribution for each feature show comparable distributions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

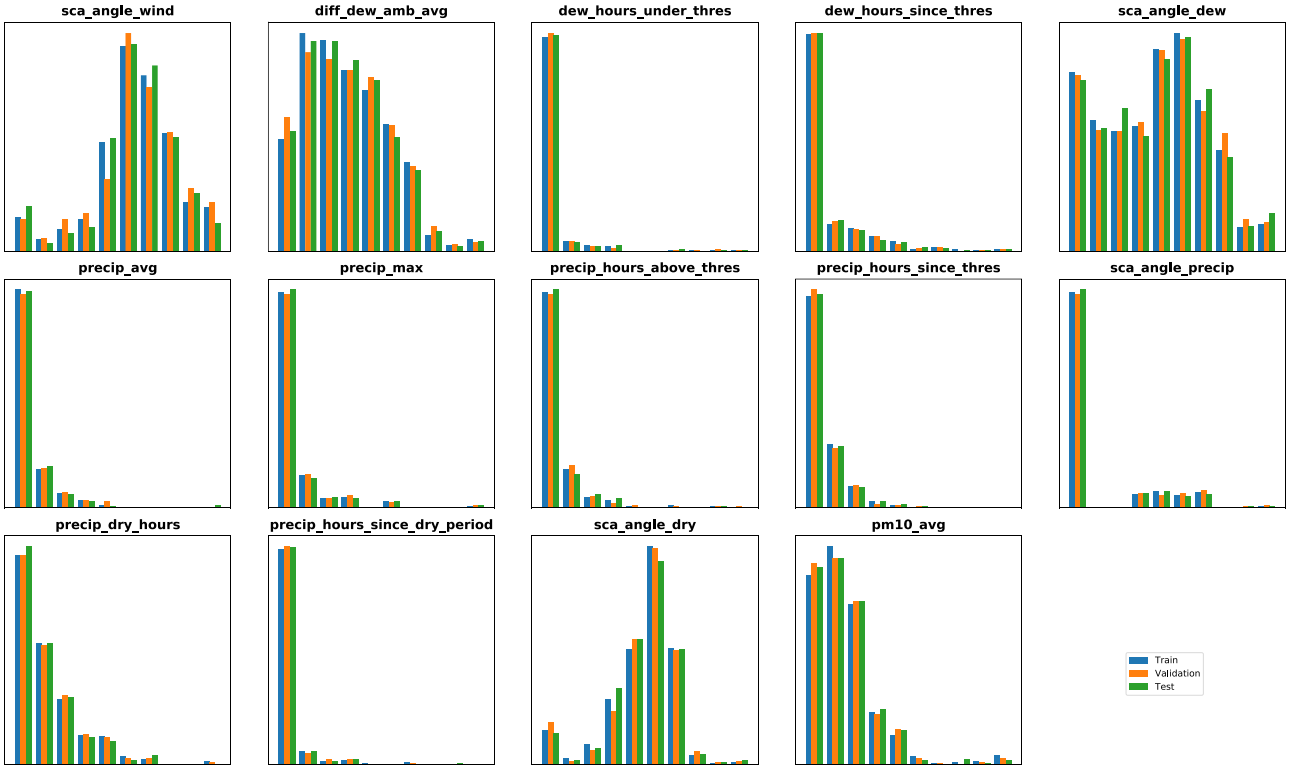

Fig. B.19. Histogram plots (part 2) of training (blue), validation (orange) and test (green) distribution for each feature show comparable distributions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

procedure (see Table B.6). The lowest MSE was achieved with 1024 neurons, but with the highest relative runtime. The second best MSE was achieved with 512 neurons, at 70% of the 1024 neuron runtime. In order to have a low MSE at a reasonable runtime, five hidden layers with 512 neurons was chosen as Neural Network architecture for the remainder of the experiments in this work.

Table B.7
Description of used features with mean min and max values and their usage in different datasets. Features are grouped into operational data, which describe features that are individual for each collector and meteorological data which are mostly identical for all collectors.

Feature	Description	mean	min	max	dataset
Operational data	Data from solar field from day of gloss measurement (average values)				
week of year sin	week of year sin = Sin(week * (π/26))	0.08	−1.00	1.00	0; 1; 2; 3
week of year cos	week of year cos = Cos(week * (π/26))	−0.19	−1.00	1.00	0; 1; 2; 3
days since last cleaning	Days between collector in cleaning position and gloss measurement	8.73	0.46	126.00	0; 1; 2; 3
collector pos x	Collector position from west to east	37.40	1	76	0; 1; 2; 3
collector pos y	Collector position from south to north	4.53	1	8	0; 1; 2; 3
collector pos loop	Collector position in Loop according to flow direction	2.50	1	4	0; 1; 2; 3
irradiance	Daily average irradiance from sunrise to sunset on collector aperture Avg(DNI * cos(phi))	462.24	13.92	847.36	0; 1; 2; 3
cumul irradiance	Cumulative sum of theoretical collected irradiance Sum(DNI * cos(phi) * IAM)	6.36E+04	910.38	1.43E+05	0; 1; 2; 3
focus	Daily average focus factor from sunrise to sunset	0.79	0.00	0.99	0; 1; 2; 3
dumping	Daily average of dumping offset	0.39	0.00	5.54	0; 1; 2; 3
collector temperature	Daily average collector temperature from sunrise to sunset [°C]	290.41	44.72	369.16	1; 3
Meteodata	Meteorological data taken from last time the collector was in cleaning position up to time of gloss measurement				
wind u avg	Wind u component average (east–west, 90° to collector axis) [$\frac{m}{s}$]	−0.83	−6.56	3.61	2; 3
wind v avg	Wind v component average (north–south, aligned with collector axis) [$\frac{m}{s}$]	0.30	−2.93	2.95	2; 3
wind hours above threshold	Sum of hours above theoretical saltation velocity (5.9 m/s at 10 m measurement height)	26.18	0.00	381.76	3
wind hours since threshold	Hours since the last time the wind was above theoretical saltation velocity	8.19	0.00	220.91	3
sca angle wind	Average collector angle during the time the collector was above saltation velocity	103.77	−4.49	181.28	3
diff dew amb avg	Average difference between the dew point temperature and ambient air temperature	9.74	2.41	24.22	2; 3
dew hours under thres	Sum of hours under dew formation threshold (5 K)	83.19	0.00	1808.64	3
dew hours since threshold	Hours since last theoretical dew formation	20.08	0.00	180.03	3
sca angle dew	Average collector angle during last theoretical dew formation	67.77	−14.16	181.38	3
precip avg	Average precipitation in mm	0.48	0.00	7.30	2; 3
precip max	Maximum precipitation in mm	3.57	0.00	50.90	3
precip hours above threshold	Sum of hours precipitation was above theoretical cleaning threshold (5 mm)	9.94	0.00	216.00	3
precip hours since threshold	Hours since precipitation was above last theoretical cleaning threshold	132.87	10.95	1436.71	3
sca angle precip	Average collector angle during last time precipitation was above theoretical cleaning threshold	18.77	−14.20	181.99	3
precip dry hours	Hours of longest dry period with no precipitation (0 mm)	95.96	0.00	600.00	3
precip hours since dry period	Hours between the last day of the dry period and gloss measurement	56.84	0.00	1364.55	3
sca angle dry	Average collector angle during dry period with no precipitation (0 mm)	87.27	−3.10	181.80	3
pm 10	Average PM10 value [$\frac{\mu g}{m^3}$]	1.57E−08	3.99E−09	6.29E−08	3

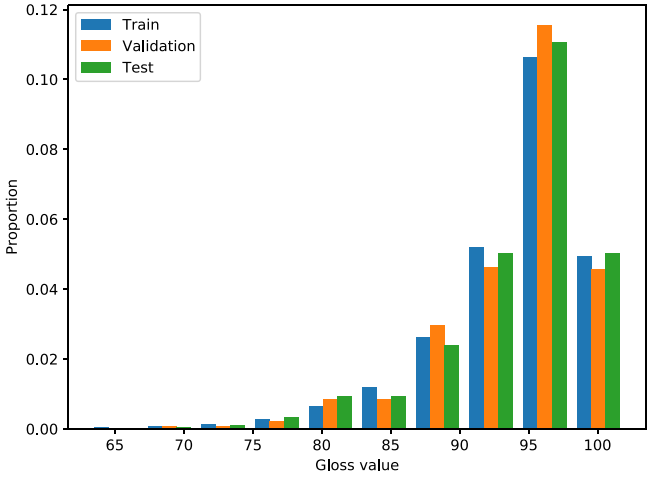

Fig. B.20. The gloss values show similar distributions across the train, validation, and test data subsets, but exhibit a clear class imbalance.

Table B.8
Model configuration parameters for Linear Regression, Decision Tree, Gaussian Process, Support Vector Machine and Neural Network. Any parameters not shown used default values.

Model	Parameters
LR	fit_intercept=True, normalize=False, copy_X=True, positive=False
DT	base_estimator=DecisionTreeRegressor(min_samples_leaf=5), n_estimators=500, learning_rate=1e−07
GP	kernel=(0.5 * RationalQuadratic(alpha=1, length_scale=1) + 1 * WhiteKernel(noise_level=1)), n_restarts_optimizer=0, normalize_y=True
SVR	kernel = 'rbf', cache_size=10000
NN	batch_size = 100, optimizer = 'Adam', loss_function=torch.nn.MSELoss()

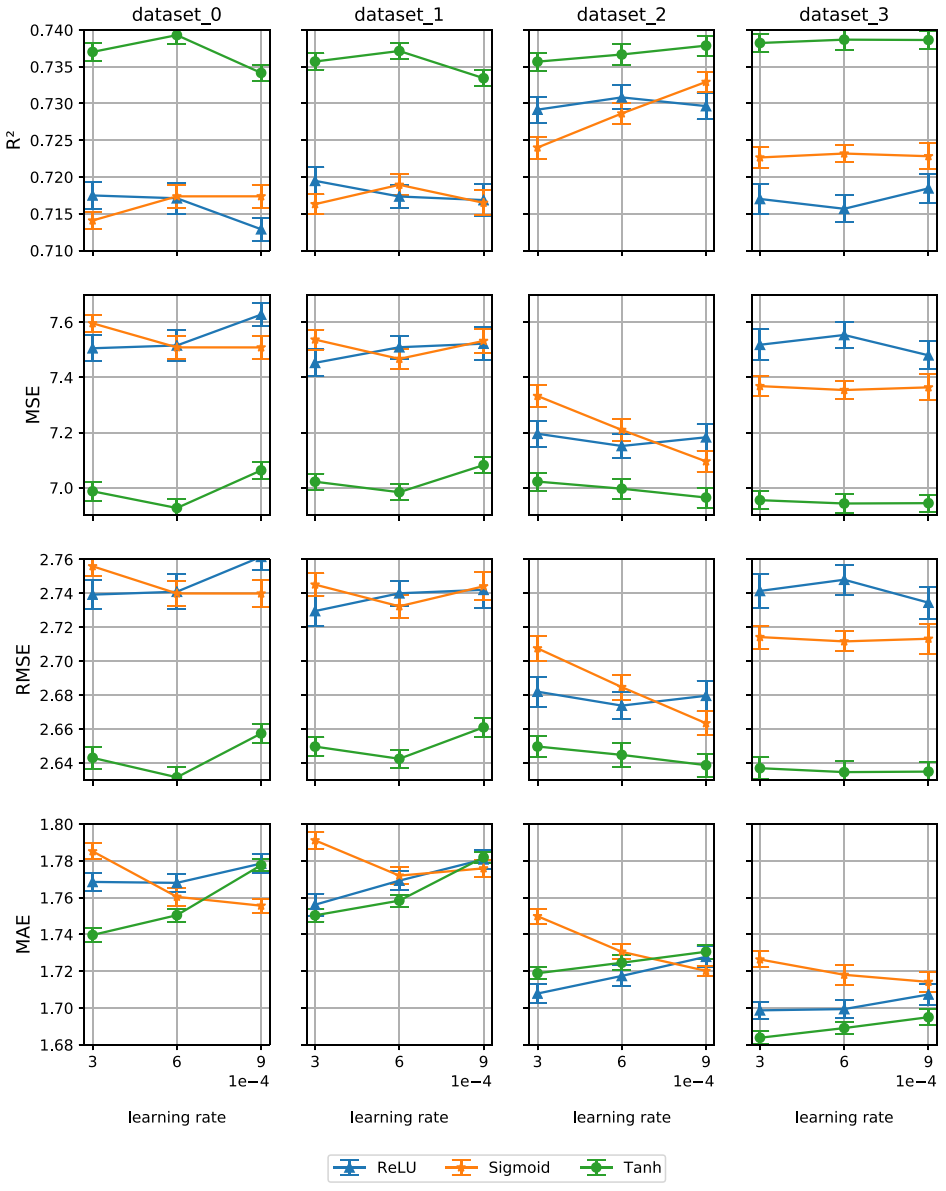

Fig. B.21. R^2, MSE, RMSE, and MAE with uncertainties from Neural Network grid search calculated from 30 different random initializations for each model configuration. The four columns represent the different datasets. Tanh activation performs best for each dataset, compared to ReLU and Sigmoid activation. Deviations between different datasets are small.

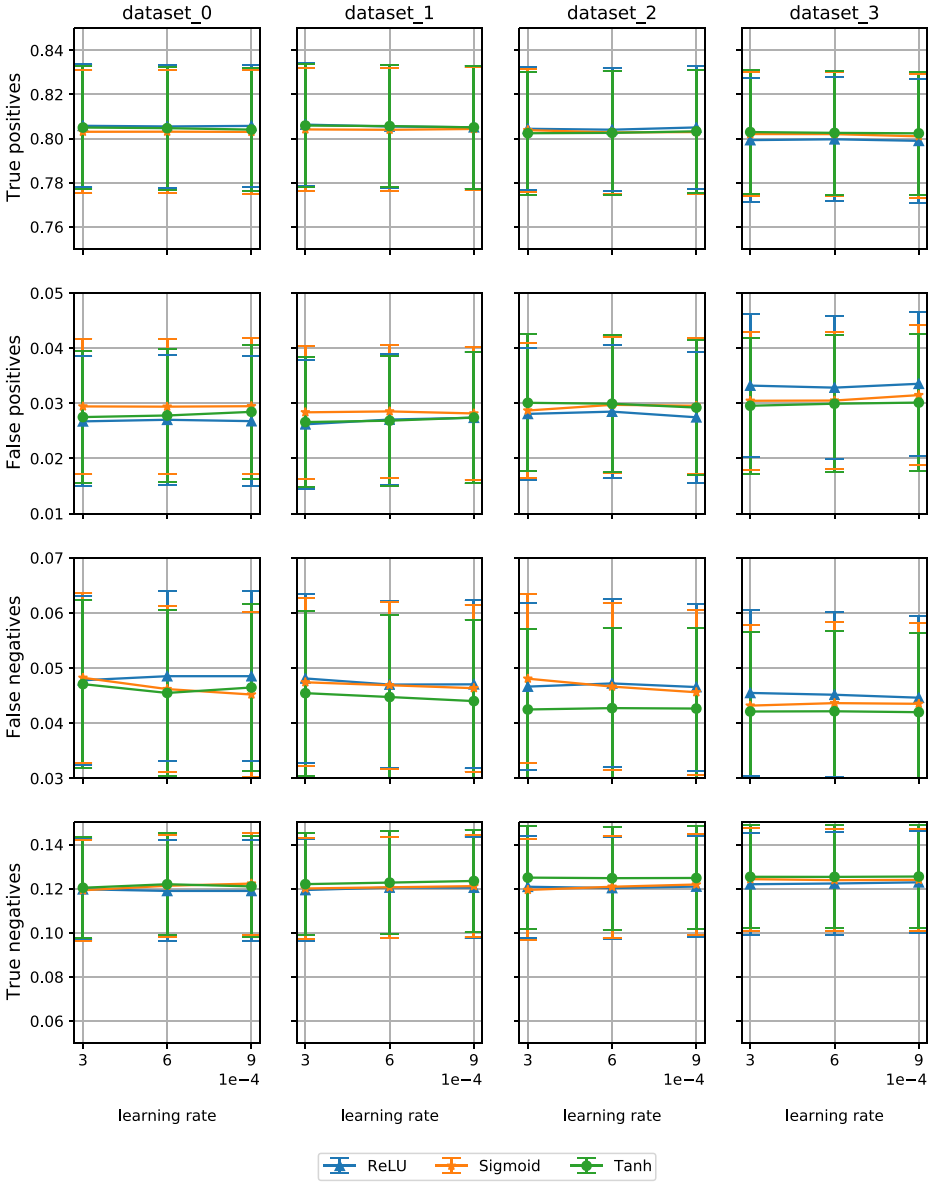

Fig. B.22. True positives, false positives, false negatives, and true negatives with uncertainties from Neural Network grid search calculated from 30 different random initializations for each model configuration. The model shows a nearly consistent performance for different datasets and different activation functions.

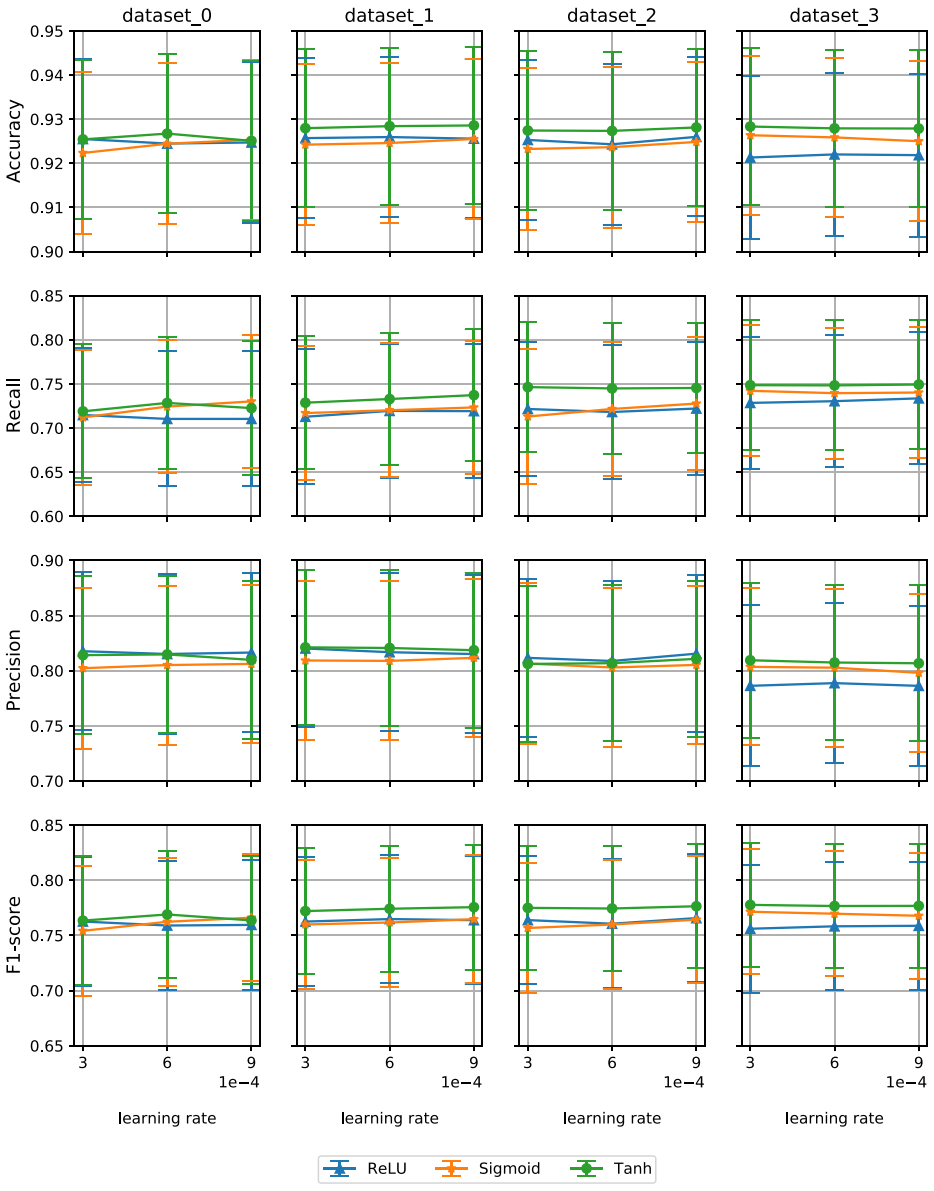

Fig. B.23. Accuracy, Recall, Precision and F1-score with uncertainties from Neural Network grid search calculated from 30 different random initializations for each model configuration. The four columns represent the different datasets. The model shows a nearly consistent performance for different datasets and different activation functions.

References

[1] K. Ilse, L. Micheli, B.W. Figgis, K. Lange, D. Daßler, H. Hanifi, F. Wolfertstetter, V. Naumann, C. Hagendorf, R. Gottschalg, J. Bagdahn, Techno-economic assessment of soiling losses and mitigation strategies for solar power generation, Joule 3 (10) (2019) 2303–2321, http://dx.doi.org/10.1016/j.joule.2019.08.019, URL https://www.sciencedirect.com/science/article/pii/S2542435119304222.

[2] P. Bellmann, F. Wolfertstetter, R. Conceição, H.G. Silva, Comparative modeling of optical soiling losses for CSP and PV energy systems, Sol. Energy 197 (2020) 229–237, http://dx.doi.org/10.1016/j.solener.2019.12.045, URL https://www.sciencedirect.com/science/article/pii/S0038092X19312599.

[3] A. Brenner, T. Hirsch, M. Röger, R. Pitz-Paal, State-of-the-art measurement instrumentation and most recent measurement techniques for parabolic trough collector fields, Energies 14 (21) (2021) 7166, http://dx.doi.org/10.3390/en14217166, URL https://www.mdpi.com/1996-1073/14/21/7166.

[4] S. Pulipaka, F. Mani, R. Kumar, Modeling of soiled PV module with neural networks and regression using particle size composition, Sol. Energy 123 (2016) 116–126, http://dx.doi.org/10.1016/j.solener.2015.11.012, URL https://www.sciencedirect.com/science/article/pii/S0038092X15006180.

[5] W. Javed, B. Guo, B. Figgis, Modeling of photovoltaic soiling loss as a function of environmental variables, Sol. Energy 157 (2017) 397–407, http://dx.doi.org/10.1016/j.solener.2017.08.046, URL https://www.sciencedirect.com/science/article/pii/S0038092X17307260.

[6] B. Laarabi, O. May Tzuc, D. Dahlioui, A. Bassam, M. Flota-Bañuelos, A. Barhdadi, Artificial neural network modeling and sensitivity analysis for soiling effects on photovoltaic panels in Morocco, Superlattices Microstruct. 127 (2019) 139–150, http://dx.doi.org/10.1016/j.spmi.2017.12.037, URL https://www.sciencedirect.com/science/article/pii/S0749603617323674.

[7] K. Chiteka, R. Arora, S.N. Sridhara, A method to predict solar photovoltaic soiling using artificial neural networks and multiple linear regression models, Energy Syst. 11 (4) (2020) 981–1002, http://dx.doi.org/10.1007/s12667-019-00348-w.

[8] N. Simal Pérez, J. Alonso-Montesinos, F.J. Batlles, Estimation of soiling losses from an experimental photovoltaic plant using artificial intelligence techniques, Appl. Sci. 11 (4) (2021) 1516, http://dx.doi.org/10.3390/app11041516, URL https://www.mdpi.com/2076-3417/11/4/1516.

[9] M. Coello, L. Boyle, Simple model for predicting time series soiling of photovoltaic panels, IEEE J. Photovolt. 9 (5) (2019) 1382–1387, http://dx.doi.org/10.1109/JPHOTOV.2019.2919628.

[10] W.F. Holmgren, C.W. Hansen, M.A. Mikofski, pvlib Python: a Python package for modeling solar energy systems, J. Open Source Softw. 3 (29) (2018) 884, http://dx.doi.org/10.21105/joss.00884.

[11] G. Picotti, P. Borghesani, G. Manzolini, M.E. Cholette, R. Wang, Development and experimental validation of a physical model for the soiling of mirrors for CSP industry applications, Sol. Energy 173 (2018) 1287–1305, http://dx.doi.org/10.1016/j.solener.2018.08.066, URL https://www.sciencedirect.com/science/article/pii/S0038092X18308363.

[12] F. Wolfertstetter, S. Wilbert, F. Terhag, N. Hanrieder, A. Fernandez-García, C. Sansom, P. King, L. Zarzalejo, A. Ghennioui, Modelling the soiling rate: Dependencies on meteorological parameters, AIP Conf. Proc. 2126 (1) (2019) 190018, http://dx.doi.org/10.1063/1.5117715, URL https://aip.scitation.org/doi/abs/10.1063/1.5117715.

[13] D. Sbarbaro, R. Peña, E. Fuentealba, Model-Based Soiling Estimation in Parabolic Solar Concentrators, AIP Publishing, 2017, 030018, http://dx.doi.org/10.1063/1.5067034.

[14] R. Conceição, H.G. Silva, M. Collares-Pereira, CSP mirror soiling characterization and modeling, Sol. Energy Mater. Sol. Cells 185 (2018) 233–239, http://dx.doi.org/10.1016/j.solmat.2018.05.035, URL http://www.sciencedirect.com/science/article/pii/S0927024818302629.

[15] F. Wolfertstetter, K. Pottler, A. Alami Merrouni, A. Mezrhab, R. Pitz-Paal, A novel method for automatic real-time monitoring of mirror soiling rates, in: SolarPACES Conference, 2012, p. 9, URL https://elib.dlr.de/77590/.

[16] H. El Gallassi, A. Alami Merrouni, M. Chourak, A. Ghennioui, The application of artificial neural network to predict cleanliness drop in CSP power plants using meteorological measurements, in: Proceedings of the 2nd International Conference on Electronic Engineering and Renewable Energy Systems, Springer Singapore, 2021, pp. 699–707, http://dx.doi.org/10.1007/978-981-15-6259-4_73.

[17] F. Wolfertstetter, R. Fonk, C. Prahl, M. Röger, S. Wilbert, J. Fernández-Reche, Airborne soiling measurements of entire solar fields with qfly, AIP Conf. Proc. 2303 (1) (2020) 100008, http://dx.doi.org/10.1063/5.0028968, URL https://aip.scitation.org/doi/abs/10.1063/5.0028968.

[18] Zehntner GmbH, ZGM 1110: Glossmeter manual, 2015, Zehntner Testing Instruments, received by email on 9 December 2020.

[19] A. Fernández-García, F. Sutter, L. Martínez-Arcos, C. Sansom, F. Wolfertstetter, C. Delord, Equipment and methods for measuring reflectance of concentrating solar reflector materials, Sol. Energy Mater. Sol. Cells 167 (2017) 28–52, http://dx.doi.org/10.1016/j.solmat.2017.03.036.

[20] K. Brooks, M.J. Schwar, Dust deposition and the soiling of glossy surfaces, Environ. Pollut. 43 (2) (1987) 129–141, http://dx.doi.org/10.1016/0269-7491(87)90071-6, URL https://www.sciencedirect.com/science/article/pii/0269749187900716.

[21] S.A. Kalogirou, Solar Energy Engineering : Processes and Systems, Elsevier Science & Technology, Saint Louis, UNITED STATES, 2014, URL http://ebookcentral.proquest.com/lib/dlr-ebooks/detail.action?docID=1517436.

[22] I. London, Encoding cyclical continuous features - 24-hour time, 2022, 9 August, 2016, URL https://ianlondon.github.io/blog/encoding-cyclical-features-24hour-time/.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, 30, Curran Associates, Inc., 2017, URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[24] G.E. Cohen, D.W. Kearney, G.J. Kolb, Final Report on the Operation and Maintenance Improvement Program for Concentrating Solar Power Plants, Report SAND99-1290; Other: ON: DE00008378 United States 10.2172/8378 Other: ON: DE00008378 SNL English, Sandia National Laboratories (SNL),, 1999, http://dx.doi.org/10.2172/8378, URL https://www.osti.gov/servlets/purl/8378.

[25] J.F. Kok, E.J.R. Parteli, T.I. Michaels, D.B. Karam, The physics of wind-blown sand and dust, Rep. Progr. Phys. 75 (10) (2012) 106901, http://dx.doi.org/10.1088/0034-4885/75/10/106901.

[26] J.R. Caron, B. Littmann, Direct monitoring of energy lost due to soiling on first solar modules in california, IEEE J. Photovolt. 3 (1) (2013) 336–340, http://dx.doi.org/10.1109/JPHOTOV.2012.2216859.

[27] M. Mehos, H. Price, R. Cable, D. Kearney, B. Kelly, G. Kolb, F. Morse, Concentrating Solar Power Best Practices Study, Report, National Renewable Energy Laboratory (NREL),, 2020, URL https://www.nrel.gov/docs/fy20osti/75763.pdf.

[28] World Meteorological Organization, Guide to Meteorological Instruments and Methods of Observation, World Meteorological Organization, 2014, URL https://library.wmo.int/doc_num.php?explnum_id=4147.

[29] L. Micheli, M. Muller, An investigation of the key parameters for predicting PV soiling losses, Prog. Photovolt., Res. Appl. 25 (4) (2017) 291–307, http://dx.doi.org/10.1002/pip.2860, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.2860.

[30] A. Kimber, L. Mitchell, S. Nogradi, H. Wenger, The effect of soiling on large grid-connected photovoltaic systems in California and the Southwest region of the United States, in: 2006 IEEE 4th World Conference on Photovoltaic Energy Conference, Vol. 2, 2006, pp. 2391–2395, http://dx.doi.org/10.1109/WCPEC.2006.279690.

[31] N. Hanrieder, S. Wilbert, F. Wolfertstetter, J. Polo, C. Alonso, L. Zarzalejo, Why natural cleaning of solar collectors cannot be described using simple rain sum thresholds, in: Solar World Congress 2021, 2021, pp. 959–969, http://dx.doi.org/10.18086/swc.2021.37.02, URL http://proceedings.ises.org/?doi=swc.2021.37.02.

[32] F. Wolfertstetter, Auswirkungen von Verschmutzung auf konzentrierende solarthermische Kraftwerke (Ph.D. thesis), RWTH Aachen, 2016, URL http://publications.rwth-aachen.de/record/706465.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830, URL http://jmlr.org/papers/v12/pedregosa11a.html.

[34] J.L. Hodges, The significance probability of the Smirnov two-sample test, Ark. Mat. 3 (5) (1958) 469–486, URL http://archive.ymsc.tsinghua.edu.cn/pacm_download/116/6944-11512_2007_Article_BF02589501.pdf.

[35] H. Drucker, Improving regressors using boosting techniques, in: Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 107–115.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. Devito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035, URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[37] F. Wolfertstetter, S. Wilbert, J. Dersch, S. Dieckmann, R. Pitz-Paal, A. Ghennioui, Integration of soiling-rate measurements and cleaning strategies in yield analysis of parabolic trough plants, J. Solar Energy Eng. 140 (2018) http://dx.doi.org/10.1115/1.4039631.

[38] R.B. Pettit, J.M. Freese, E.P. Roth, Studies of dust accumulation on solar-mirror materials, 1981, URL https://www.osti.gov/biblio/6636866.

[39] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.

[40] C.J. Clopper, E.S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial, Biometrika 26 (4) (1934) 404–413, http://dx.doi.org/10.1093/biomet/26.4.404.

[41] H. Truong Ba, M.E. Cholette, R. Wang, P. Borghesani, L. Ma, T.A. Steinberg, Optimal condition-based cleaning of solar power collectors, Sol. Energy 157 (2017) 762–777, http://dx.doi.org/10.1016/j.solener.2017.08.076, URL https://www.sciencedirect.com/science/article/pii/S0038092X17307582.

[42] H. Truong-Ba, M.E. Cholette, G. Picotti, T.A. Steinberg, G. Manzolini, Sectorial reflectance-based cleaning policy of heliostats for solar tower power plants, Renew. Energy 166 (2020) 176–189, http://dx.doi.org/10.1016/j.renene.2020.11.129.

[43] F. Terhag, F. Wolfertstetter, S. Wilbert, T. Hirsch, O. Schaudt, Optimization of cleaning strategies based on ANN algorithms assessing the benefit of soiling rate forecasts, 2019, 220005, http://dx.doi.org/10.1063/1.5117764, URL https://aip.scitation.org/doi/abs/10.1063/1.5117764.

[44] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on tabular data? 2022, http://dx.doi.org/10.48550/ARXIV.2207.08815, URL https://arxiv.org/abs/2207.08815, preprint.

[45] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE PP (2020) 1–34, http://dx.doi.org/10.1109/JPROC.2020.3004555.

[46] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.