

## Inappropriate machine learning application in real power industry cases

Alexandra Khalyasmaa<sup>1,2</sup>, Pavel Matrenin<sup>3</sup>, Stanislav Eroshenko<sup>1,2</sup>

<sup>1</sup>Electrical Engineering Department, Ural Federal University, Ekaterinburg, Russia

<sup>2</sup>Power Plants Department, Novosibirsk State Technical University, Novosibirsk, Russia

<sup>3</sup>Industrial Power Supply Systems Department, Novosibirsk State Technical University, Novosibirsk, Russia

### Article Info

#### Article history:

Received May 3, 2021

Revised Nov 9, 2021

Accepted Nov 30, 2021

#### Keywords:

Digital transformation

Intelligent system

Machine learning application

Power generation forecasting

photovoltaic power plants

### ABSTRACT

Global digital transformation of the energy sector has led to the emergence of multiple digital platform solutions, the implementation of which have revealed new problems associated with continuous growth of data volumes requiring new approaches to their processing and analysis. This article is devoted to the improper application of machine learning approaches and flawed interpretation of their output at various stages of decision support systems development: data collection; model development, training and testing as well as industrial implementation. As a real industrial case study, the article examines the power generation forecasting problem of photovoltaic power plants. The authors supplement the revealed problems with the corresponding recommendation for industrial specialists and software developers.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



### Corresponding Author:

Alexandra Khalyasmaa

Electrical Engineering Department, Ural Federal University

19 Mira Street, Ekaterinburg, 620002, Russia

Email: a.i.khaliasmaa@urfu.ru

## 1. INTRODUCTION

There are many publications describing machine learning (ML) algorithms and their concepts, and even their specific industry applications, including those for power engineering problems [1]–[3]. However, from the point of view of data science, it is the processed data obtained as a result of filtration and other transformations as well as the data obtained as a result of data mining that represents intellectual property and, most often, is classified as commercial information. In such circumstances, it is obvious that in each specific case, the methods and approaches implemented by the authors of the study are impossible to be repeated for verification without having access to the dataset. At the same time, the features of each specific problem and its effective solution using ML algorithms almost completely depend on the dataset used to solve the problem [4]–[6].

Nowadays ML is generally recognized as an effective tool for data processing, but its correct application by the developers is a hot-button issue. The primary barrier for intelligent systems development in the power industry globally can be described as follows: modern high-end professionals in the power industry do not fully possess the required competencies in data science, and high-end information technology (IT) specialists do not fully understand the physics of power generation, transmission, and consumption. Hence, two types of global errors do appear: in the first case it is an incorrect implementation of the mathematical apparatus and software architecture; in the second case it is an improperly addressed knowledge base and an incorrect interpretation of the obtained results. Therefore, the effective implementation of such projects is feasible only by a joint team of data science and IT specialists and power

engineers, each of whom additionally possesses at least basic background in data science and in power engineering, correspondingly.

Many scientists in the sphere of power and energy do not have deep knowledge of data science and, what is also particularly important, are usually not professional programmers. Therefore, they mistakenly believe that the implementation of intelligent systems lies only in using a specific ML algorithm. If applying ML algorithms correctly, then the result is a process that combines low-level details and high-level software structure. In other words, whatever problem in the sphere of power industry you solve by ML algorithms, you ultimately create software. It is very important to be aware of this, since all the relevant stages and problems of software implementation for such tasks are inherent. Therefore, in this paper possible errors in creating automated systems based on ML are conditionally divided depending on the stages of software implementation: i) errors at the stage of data collection, analysis, and preparation; ii) modeling and testing errors; and iii) full-scale commercial implementation errors.

## **2. ERRORS AT THE STAGE OF DATA COLLECTION, ANALYSIS AND PREPARATION**

Errors at the stage of data collection, analysis, and preparation, according to the authors' viewpoint, have one of the most severe consequences for the system operation, since it is the dataset that is the basis for intelligent system, and it is faulty understanding can lead to errors in the transformation and interpretation of the results. Given that the transformed data at one stage can be used as input in another one, it becomes clear that even a small error at an early stage can multiply and completely skew the results, and lead either to a result with low accuracy, or to a completely incorrect interpretation. The main stages at which developers typically make the errors include the following: i) data sources selection, ii) data preprocessing, and iii) sampling principles.

### **2.1. Data sources selection**

Irrelevant data sources may be selected so that the data may be initially incorrect. At this stage, it is important to understand that the data sources selection entirely depends on the person at the stage of designing an automated system, therefore it is of crucial importance that such systems be developed jointly by data science specialists and power engineers. Such errors can lead to false correlations and parameters dependences, which in fact may not exist at all [6]. For example, when data on defects of dry-type power transformers of the same voltage rating are used to identify oil-filled power transformers defects. The main task of ML algorithms is to generalize data, so the machine performs data mining strictly for the dataset that the developer has chosen.

An important aspect in this situation is the initial adequacy of the dataset. Here we are not talking about outliers and occasional errors in the dataset, but rather about cases of bad ("poisoned") data, when, for example, the data represents initially defective (for example, in case of manufacturer's defect) power transformers within the same initial data sample, that is such transformers form a whole cluster. Also, the appearance of "poisoned" data can be intentional, for example, as a result of cyber-attacks, which is also an important issue nowadays. Therefore, for automated systems operating at strategic high-voltage facilities, such as power plants and substations, it is also necessary to ensure the data exchange security. The result of errors at this stage, subject to the "poisoned" data prevalence, may be completely incorrect operation of the system and inadequate data generalization by implemented models.

### **2.2. Data preprocessing**

Data preprocessing is essential for applying ML algorithms and may include the following procedures: i) feature extraction, ii) feature transformation, iii) feature interaction analysis, iv) gap filling, and v) filtration. It is the software developer who manages the data processing goals and sequencing at the system design stage as well as the developed solutions validation at the development and testing stages. When selecting relevant data sources, the lack of their preprocessing is more likely to lead to low accuracy of the developed model and low speed of operation of such a system than to systematic errors (provided that we are not pursuing real time operation of the system). For example, within the author's research, it was revealed that the absence of data preprocessing stage, on average, reduces the accuracy of the photovoltaic power plants generation forecast by 20-25% [7].

### **2.3. Sampling principles**

Another important step at the stage of data collection, analysis and preparation is selecting the principle of dividing the data into training, validation and testing sets [8], [9]. It is generally accepted that the accuracy of the algorithm operation largely depends on the training set volume. Such a statement is not always correct, since a large volume of the training set does not yet guarantee a balance within it. For

example, for the classification problem, the imbalance between the classes (the lack of certain classes data or multiple prevalence of the number of samples of one class over another one) can ultimately degrade the entire operation of the system, since the algorithm will not be able to generalize correctly. Similar problems are associated with the testing and validation sets as well. Such problems can generally be solved either by using a normalization procedure, or by adding or cutting training data.

An equally important problem is the class imbalance problem [10]. In the power industry imbalance in the training and testing set is almost always associated with the problem of detecting defects in high-voltage power equipment, regardless of the equipment type. Obviously, in the dataset in such tasks, the parameters characterizing the equipment defect-free (normal) state will prevail, or, in the worst case, certain types of defects may be completely absent. If we form the training and testing sets in accordance with the generally accepted statement that the probability of a certain type of defect in the training set is equal to the probability of the appearance of these defects in the dataset, this will lead to the fact that the system will perfectly recognize a defect-free state and rare defects most likely will be considered as “outliers” in measurements. Thus, selecting the principles of dividing dataset into training, testing and validation sets should be a separate task for the intelligent systems developer.

### 3. MODELING AND TESTING ERRORS

#### 3.1. General errors

One of the fundamental errors in applying ML algorithms for a specific problem in the power industry is the lack of justification for their application. Despite the effectiveness of this mathematical apparatus, the developers of intelligent systems must first make sure of the real necessity to use ML algorithms, namely, to clearly define the problem category in terms of its mathematical formulation, the data sufficiency for its correct implementation, and also make clear justification that the use of traditional analytical deterministic approaches to data analysis provides poor results.

ML algorithms are usually worth using in problems with so-called big data. But there are ML algorithms that can really be effective for a small amount of data, but for each specific problem and each individual algorithm, it is necessary to additionally determine the minimum required and sufficient amount of data to assure correct generalizing ability of the algorithm. One of the main challenges in modeling intelligent systems in the energy sector using ML algorithms is the correct formulation of the ML problem and its categorization:

- Regression problem determining (forecasting) a continuous dependent variable (or several variables) from a number of independent variables (for example, forecasting power plants generation or power consumption) [11].
- Classification problem dividing objects according to previously known classes (for example, the technical state analysis of power equipment according to the indicators of its operation) [12], [13].
- Clustering problem dividing objects into groups (clusters) depending on their similarity, provided that the list of clusters is not clearly specified in advance and it is determined during the algorithm operation, including that one of the subtasks is to determine the intra-clusters relations (for example, identification of various types of defects in high-voltage power equipment based on various data of technical diagnostics) [14], [15].

The regression problem, like the classification one, is supervised learning problem, which is solved based on preliminary labeled data. The clustering problem is an unsupervised learning problem. Each of the categories has its own characteristics, application sphere, advantages, and disadvantages. Often, researchers use a simple enumeration of methods in finding solution for the problems being analyzed. And in the most cases this enumeration is based on the expert opinion and developers’ personal experience, and the justification for the necessity and feasibility of certain algorithms application does not always look convincing.

It is also worthwhile to divide problems according to the required time to solve them and the required time to train the ML model. An operational task that requires a large (predetermined) high quality data and a short training time for the model and assumes online operation of the system or at least near the real-time. For example, the task of power balance operational planning in power systems [16], determining the required active power reserves, taking into account the probability of power imbalance, where the load, power generation for the selected lead intervals, operational forecast of renewable energy sources, power grid constraints are taken as initial data. The solution of such problems in their industrial implementation is always associated with the need to build the corresponding infrastructure: the data warehouses and powerful distributed computing facilities.

A medium-term task that requires enough data to get a good accuracy result in a reasonable time. For example, the tasks of diagnosing the technical state of power equipment in order to identify developing

defects [17], where the initial data contains the results of technical diagnostics of the power equipment unit and its elements in case there is no preinstalled technical state monitoring system.

A long-term task, the main requirements of which are to increase the accuracy while reducing the training time under limited data conditions (either a small amount of data, or a large amount of data of insufficient quality). For example, the task of designing a strategic development plan of the national power system in order to ensure reliable power supply [18], [19], where the general retrospective data on the power system, load, generation, data on the expected power grid development, changes in power consumption, long-term balances of power and electric power are included into the initial dataset. Usually this is the so-called class of advice-giving systems or decision support systems.

### 3.2. Particular errors

Particular errors usually result from the lack of mathematical or software developing experience of the developers. Most often, errors are associated with the following stages: i) selecting the way of training the model (supervised, unsupervised, with reinforcement), ii) selecting performance (quality) criteria for the model (metrics selection), iii) errors' analysis resulting from the algorithms operation and their interpretation, and iv) adaptation of the system in the case of new objects appearance.

## 4. ERRORNEOUS INDUSTRY CASES FOR POWER GENERATION FORECASTING PROBLEMS

In industrial operation of decision support systems based on ML algorithms, the above errors can occur both individually and all at once. Within this section, the authors provide the possible errors analysis and their influence on the system operation results as in the case of power generation forecasting of a photovoltaic power plant. The Figures 1-4 are just illustrations, obtained from synthetic data.

Photovoltaic power plants demonstrate the highest dynamics of growth among renewable-based power plants worldwide [7], [20]. The relevance of the problem being solved has no doubts. The need to forecast the power generation of renewable energy sources is recognized internationally [21]. Various ML methods are applied for this problem: artificial neural networks [22]–[24], population-based algorithms [25], support vector regression [26], [27], regression trees, and ensembles of regression tree [27], [28]. However, at the time of this writing, the authors do not know an internationally recognized reliable industry solution to the problem of photovoltaic power plants generation forecasting, implemented into the process-related activities of the key power industry stakeholders world-wide. Aiming to increase the efficiency of power system short-term planning in terms of compliance with system constraints and active power reserves requirements, short-term (day-ahead) photovoltaic power plants' generation forecast is formed. Research objective is to develop a model of a day-ahead (short-term) photovoltaic power plants' generation forecasting system.

### 4.1. Errors at the data collection stage

In rare cases, to solve the problem of photovoltaic power plants' generation forecasting, as for any other problem in real life, there is a ready-made dataset a set of processed data suitable for processing by ML algorithms. The formation of such a dataset is not just a task of collecting data, but also, what is very important, ranking their sources by relevance, where relevance means the relation (correspondence) degree of the analyzed object in the dataset to your specific task. For example, if we exclude the process of data sources ranking in the photovoltaic power plants' generation forecasting problem, then the dataset may contain data that is irrelevant for the problem being solved, for example: i) data from power plants located in different climatic zones or data collected only at certain times of the year, which will result in missing the trend and/or seasonal component in the time series, and ii) data from photovoltaic power plants that differ significantly by the types of the solar panels, by the types of other equipment, by the switchgear configuration, which will increase the variance of the predicted value.

Thus, it is impossible to randomly form the initial dataset; data ranking must be implemented in respect to various factors that confirm the relevance of the data sources. As a particular case, let's consider the following scenario. The customer the company owning a number of photovoltaic power plants titled from *A* to *K* (11 objects) sets out to develop a system for their generation schedules forecasting. For illustration purposes, Figure 1(a) shows the layout of a complete dataset from these power plants in the axis's "month" "latitude" (these features were selected primarily for greater clarity and simplicity of the figure).

The crosses mark the data that was included to the dataset due to a poor planning of the data collection stage. It can be seen that the set contains data for each month, but at the same time there is not any power plant for which the set would contain data for all months of the year. A certain range of latitudes (climatic zones) is covered, but there is no data for plants *B* and *C* within this range. As a result, at the stage

of developing, evaluating, and testing the model (we assume that the set is randomly divided into training and testing samples), the results on the testing set may turn out to be good enough for the task at hand, but at the same time, the model will not be suitable for application in practice. The model accuracy for the entire system is shown in Figure 1(b), the  $R^2$  metrics is taken as an accuracy criterion.

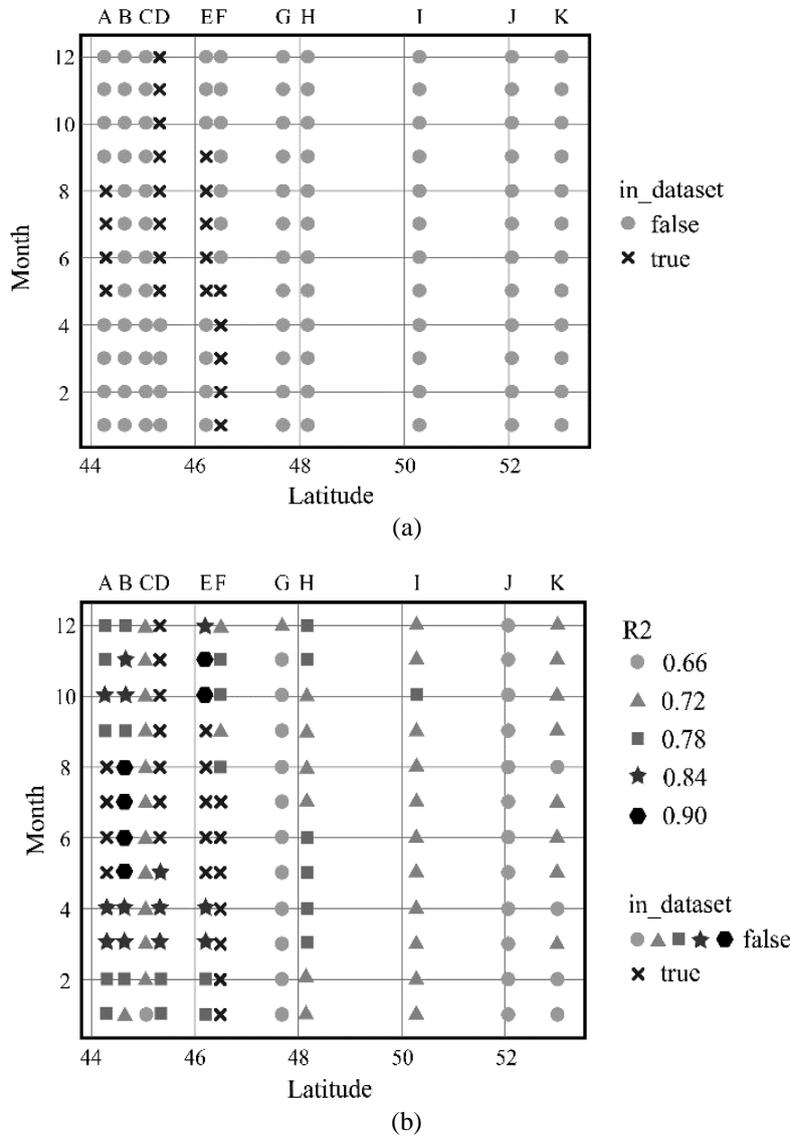


Figure 1. Initial dataset for (a) photovoltaic power plants forecasting and (b) the forecasting accuracy across the entire set of photovoltaic power plants

The decrease in forecasting accuracy for power plants *G* and *K* occurs for obvious reasons the set did not contain data from power plants located at these latitudes. But due to the time intervals (months) were different for different power plants, it turned out that, despite the presence of data in the set for the entire year, the developed model cannot be applied to any of the power plants throughout the year. The power plant *E* is an exception, since the set contained data from the power plants *D* and *F* located close to it, covering all months of the year. In addition, the results for power plant *C* were unexpectedly low, notably in any of the months. This occurs is due to the power plant *C* specifics (power equipment, solar panels type and power plants' grid connection configuration). As a result, at the testing stage, the accuracy of the model was high, while the testing set contained data for different months of different power plants located at different latitudes. But at the implementation stage, it turns out that the model does not operate properly not only for photovoltaic power plants at other latitudes, but in other weather conditions, which could be expected, and for the power plants of a different type located in the same latitudes. And the worst thing is that the model

demonstrates low accuracy in some months, even for the very power plants from which the original data was collected.

Therefore, at the stage of data collection, it is necessary: i) to clearly define the conditions in which the model is expected to operate, and coordinate them with the customer; ii) to collect data so that the dataset contains all the required conditions in sufficient volume and quality for training and testing; and iii) to understand that not only all conditions should be presented, but also combinations of conditions, in case of their correlation. In the example considered above, it would be necessary to start with collecting data for all months and even several years of one photovoltaic power plant, and only after successfully checking the adequacy of the model after its implementation and commissioning, to proceed to scaling the model (adaptation, retraining, or even building a completely new model) for other plants.

#### 4.2. Errors at the data preprocessing stage: the outliers

In the previous section of this paper, the authors described the importance of the data preprocessing stage. Below is an example of a regression model training that forecasts the photovoltaic power plant generation. Figure 2 shows a fragment of the power generation curve before preprocessing (it contains large errors highlighted by circles) and after (large errors are eliminated from the dataset). In the general case, such errors can be associated with measurement system malfunctions, data transfer failures, errors in the program code when converting and readings files, typos made during manual copying. The large errors shown in Figure 2 were intentionally integrated into the dataset by hand to demonstrate this example.

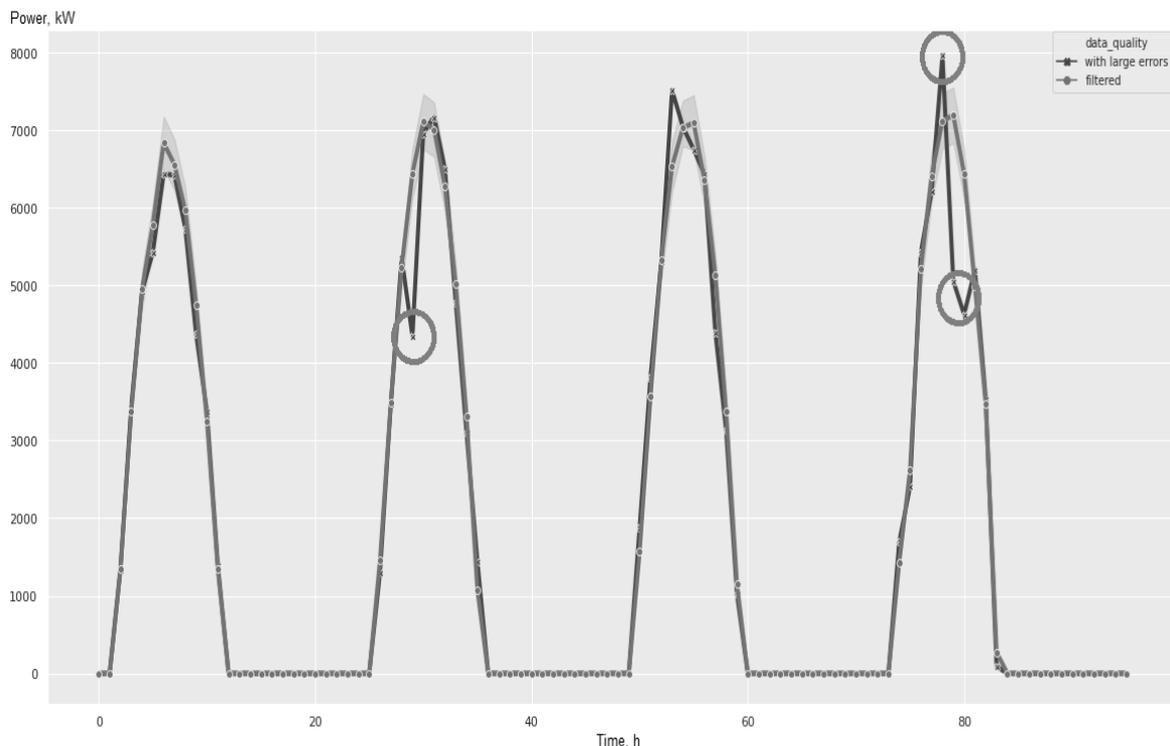


Figure 2. The power plant generation curve with large errors and after their removal

During training stage, the model strives to find dependencies in the initial dataset (energy generation schedule, date and time, meteorological data: temperature, cloudiness, humidity, wind speed). At the same time, large errors in the dataset can corrupt true dependencies and even lead to the detection of false ones. If the tools to control model overfitting are used, then it is possible to avoid false dependencies. Nevertheless, large errors reduce the accuracy, since they distort the values of the loss function and thus worsen the training process convergence. In the case of false dependencies, the model may give a forecast with large errors in certain hours of solar power plant operation, as shown in Figure 3. In this case, such errors will be for the user evidence that the model contains wrong rules, which will significantly reduce the credibility to the forecasts of the model, which is already a “black box” due to the ML algorithms application.

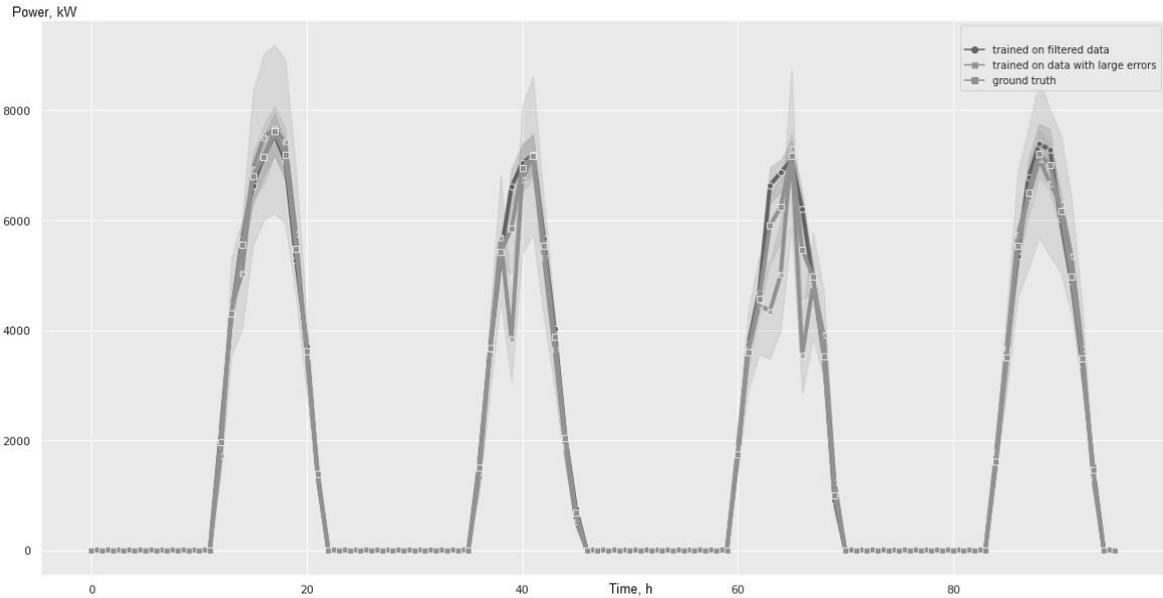


Figure 3. The output of models trained on erroneous and filtered data

**4.3. The error of selecting the irrelevant quality metrics**

The model accuracy factor selection determines how the training process takes place and how the model results are analyzed. At the same time, both in classification and regression problems there is a large number of different quality metrics. Selection of the irrelevant quality metrics can be misleading since it may not reflect the model quality in terms of its operational efficiency. Figure 4 shows a daily curve of solar power plant generation, a forecast produced by the model, and two graphs of two different quality metrics: the absolute percentage error  $|(y-y^*)/y|$  and the absolute error  $|y-y^*|$ . In this example, when the error is inherently large, the percentage error values are low due to the large actual values of solar power plant generation. On the other hand, in the boundary hours (morning and evening), an insignificant absolute error leads to huge percentage error.

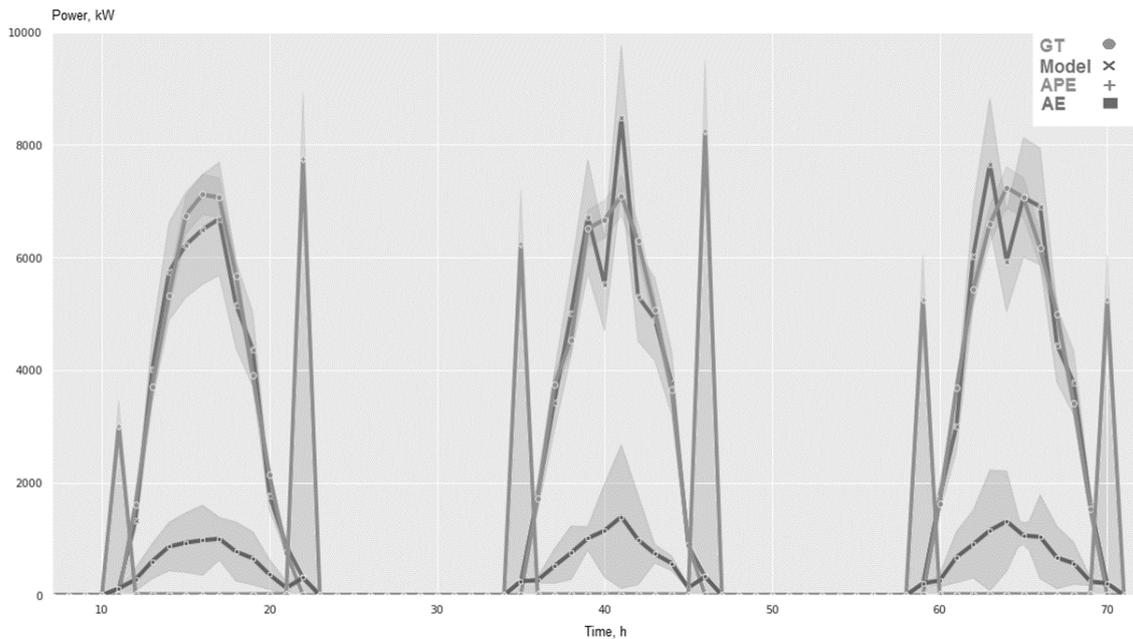


Figure 4. The model quality assessment: ground truth (GT), model the output of the regression model, absolute percentage error (APE), and absolute error (AE)

If we use the percentage error when training the model, then the model will always tend to give zero output, because even small deviations in the boundary hours will give very large values of the percentage error (more than 100%). Therefore, it is necessary: i) to select and interpret the model quality metrics based on the peculiar features of the problem under consideration; and ii) to understand what exactly lies behind the selected quality metrics, not according to mathematical laws, but according to physical and economic criteria.

#### 4.4. The error of selecting a model unsuitable for a particular problem

Selecting a model and ML algorithm, which by their characteristics do not correspond to the task under consideration, will lead to the model accuracy degradation. Such errors are less dangerous than the errors discussed above, since their negative effect reveals immediately in the course of the model training phase, not at the operational stage. Nevertheless, attempts to use inappropriate models can significantly increase the labor costs for intelligent system development, and in the worst case, result in the conclusion that it is impossible to achieve the required quality indicators.

As an example, the application of the following models to the same problem of solar power plant generation forecasting is considered: polynomial regression; decision tree; gradient boosting on decision trees. Obviously, in the case study under consideration, the model should be able to predict fundamentally different daytime hours, including daylight time and the period after sunset and before dawn. The polynomial regression model is not capable to deal with such logic since it is inherently a continuous function. The decision tree, on the contrary, forms a piecewise continuous function and can easily learn the logic of separating the operation and non-operation intervals, but with low depth it will not be able to accurately forecast the generation due to its discreteness. To solve the problem, a shallow decision trees ensemble can be effectively applied. The results of the models for this fragment are shown in Figure 5. Polynomial regression cannot learn to cut off the night hours and therefore goes negative. A shallow decision tree gives too stepped output since it cannot accurately describe the graph due to its discreteness. To avoid the erroneous selection of an inappropriate model it is necessary: i) to understand the problem features and the nature of target value changes; ii) to know the mathematical nature of the models and their learning principles; and iii) not only to analyze the model quality metrics, but also to compare its output with the required one.

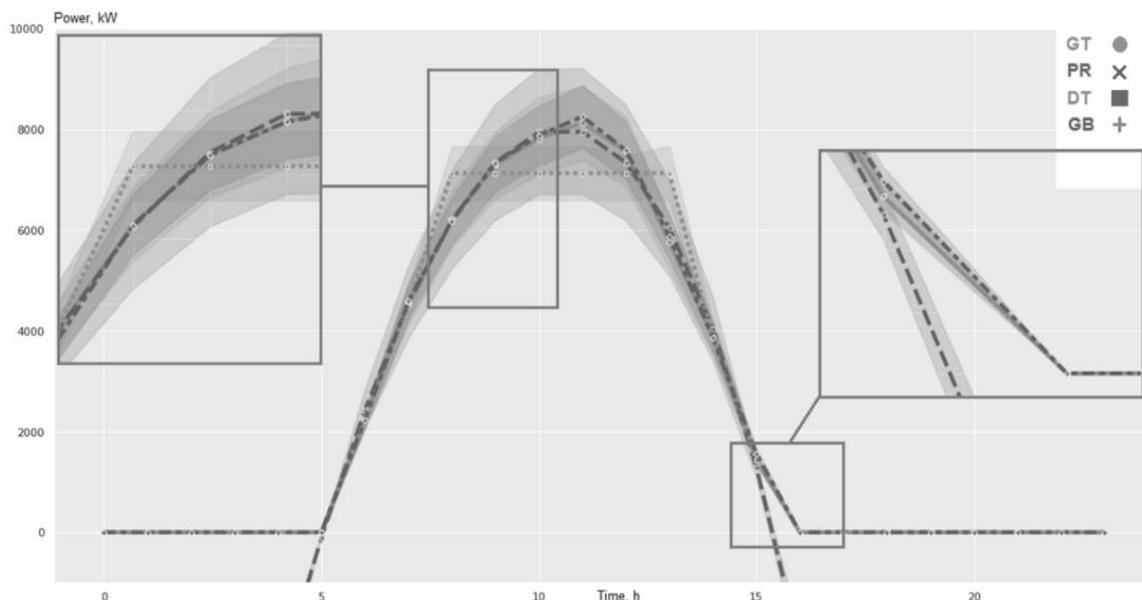


Figure 5. Regression model results: GT, polynomial regression (PR), decision tree (DT), and gradient boosting (GB)

## 5. CONCLUSION

The main types of errors that are allowed when building intelligent systems using ML are considered. The reasons and consequences of some errors are shown on the example of power generation forecasting. The considered cases of “poisoned” intelligent system substantiate the absolute need for close

interaction of specialists in data science with specialists in the power industry at all stages of intelligent system development. At each stage, it is necessary to have a complete understanding of the specifics of the problem being solved and the object involved, as well as a keen understanding of the principles of mathematical models, algorithms, and statistics. The lack of the first or the latter one with a high probability will lead either to dramatic growth of labor costs for the intelligent system development or to the model performance index degradation at the stage of its industrial operation.

## ACKNOWLEDGEMENTS

The reported study was supported by Russian Foundation for Basic Research RFBR, research project No. 20-010-00911.

## REFERENCES

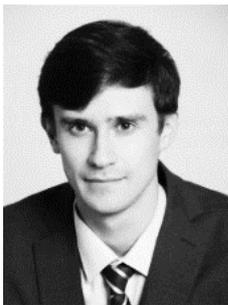
- [1] R. J. Bessa, "Future trends for big data application in power systems," in *Big Data Application in Power Systems*, Elsevier, 2018, pp. 223–242.
- [2] Y. Zhang, S. Ma, H. Yang, J. Lv, and Y. Liu, "A big data driven analytical framework for energy-intensive manufacturing industries," *Journal of Cleaner Production*, vol. 197, pp. 57–72, Oct. 2018, doi: 10.1016/j.jclepro.2018.06.170.
- [3] S. R. Salkuti, "A survey of big data and machine learning," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 575–580, Feb. 2020, doi: 10.11591/ijece.v10i1.pp575-580.
- [4] O. A. Alimi, K. Ouahada, and A. M. Abu-Mahfouz, "A review of machine learning approaches to power system security and stability," *IEEE Access*, vol. 8, pp. 113512–113531, 2020, doi: 10.1109/ACCESS.2020.3003568.
- [5] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, "Do we need more training data?," *International Journal of Computer Vision*, vol. 119, no. 1, pp. 76–92, Aug. 2016, doi: 10.1007/s11263-015-0812-2.
- [6] A. I. Khalyasmaa, M. D. Senyuk, and S. A. Eroshenko, "Analysis of the state of high-voltage current transformers based on gradient boosting on decision trees," *IEEE Transactions on Power Delivery*, vol. 36, no. 4, pp. 2154–2163, Aug. 2021, doi: 10.1109/TPWRD.2020.3021702.
- [7] S. A. Eroshenko, A. I. Khalyasmaa, D. A. Snegirev, V. V. Dubailova, A. M. Romanov, and D. N. Butusov, "The impact of data filtration on the accuracy of multiple time-domain forecasting for photovoltaic power plants generation," *Applied Sciences*, vol. 10, no. 22, Nov. 2020, doi: 10.3390/app10228265.
- [8] I. L. Kaftannikov and A. V. Parasich, "Problems of training set's formation in machine learning tasks," *Bulletin of the South Ural State University*, vol. 16, no. 3, pp. 15–24, 2016, doi: 10.14529/ctcr160302.
- [9] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials*, vol. 5, no. 1, Dec. 2019, doi: 10.1038/s41524-019-0221-0.
- [10] Y. Jian, M. Ye, Y. Min, L. Tian, and G. Wang, "FORF-S: a novel classification technique for class imbalance problem," *IEEE Access*, vol. 8, pp. 218720–218728, 2020, doi: 10.1109/ACCESS.2020.3040978.
- [11] E. D. Obando, S. X. Carvajal, and J. Pineda Agudelo, "Solar radiation prediction using machine learning techniques: a review," *IEEE Latin America Transactions*, vol. 17, no. 4, pp. 684–697, Apr. 2019, doi: 10.1109/TLA.2019.8891934.
- [12] P. Mirowski and Y. LeCun, "Statistical machine learning and dissolved gas analysis: a review," *IEEE Transactions on Power Delivery*, vol. 27, no. 4, pp. 1791–1799, Oct. 2012, doi: 10.1109/TPWRD.2012.2197868.
- [13] Y. Benmahamed, M. Tegar, and A. Boubakeur, "Application of SVM and KNN to duval pentagon 1 for transformer oil diagnosis," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 24, no. 6, pp. 3443–3451, Dec. 2017, doi: 10.1109/TDEI.2017.006841.
- [14] V. Tra, B.-P. Duong, and J.-M. Kim, "Improving diagnostic performance of a power transformer using an adaptive over-sampling method for imbalanced data," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 26, no. 4, pp. 1325–1333, Aug. 2019, doi: 10.1109/TDEI.2019.008034.
- [15] X. Hao, C. Tao, L. Rui-jing, L. Jian, and S. Cai-xin, "Fault diagnosis of power transformer using kernel-based possibilistic clustering," in *2006 International Conference on Power System Technology*, Oct. 2006, pp. 1–5, doi: 10.1109/ICPST.2006.321491.
- [16] J. Hu, X. Wei, M. Yang, B. Tang, K. Lin, and Y. Zhong, "A practical RBF framework for database load balancing prediction," in *2020 3rd International Conference on Artificial Intelligence and Big Data, ICAIBD 2020*, May 2020, pp. 83–86, doi: 10.1109/ICAIBD49809.2020.9137481.
- [17] M. Dong, W. Li, and A. B. Nassif, "Long-term health index prediction for power asset classes based on sequence learning," *IEEE Transactions on Power Delivery*, p. 1, 2021, doi: 10.1109/TPWRD.2021.3055622.
- [18] M. Glavic, R. Fonteneau, and D. Ernst, "Reinforcement learning for electric power system decision and control: past considerations and perspectives," in *IFAC*, Jul. 2017, vol. 50, no. 1, pp. 6918–6927, doi: 10.1016/j.ifacol.2017.08.1217.
- [19] R. Donida Labati, A. Genovese, V. Piuri, F. Scotti, and G. Sforza, "A decision support system for wind power production," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 1, pp. 290–304, Jan. 2020, doi: 10.1109/TSMC.2017.2783681.
- [20] "Renewables 2017 global status report," 2017. Accessed: Feb. 22, 2021. [Online]. Available: [https://www.ren21.net/wp-content/uploads/2019/05/GSR2017\\_Full-Report\\_English.pdf](https://www.ren21.net/wp-content/uploads/2019/05/GSR2017_Full-Report_English.pdf).
- [21] A. van Stiphout, T. Brijs, R. Belmans, and G. Deconinck, "Quantifying the importance of power system operation constraints in power system planning models: a case study for electricity storage," *Journal of Energy Storage*, vol. 13, pp. 344–358, Oct. 2017, doi: 10.1016/j.est.2017.07.003.
- [22] I. Khan, H. Zhu, J. Yao, D. Khan, and T. Iqbal, "Hybrid power forecasting model for photovoltaic plants based on neural network with air quality index," *International Journal of Photoenergy*, vol. 2017, pp. 1–9, 2017, doi: 10.1155/2017/6938713.
- [23] A. Chaouachi, R. M. Kamel, R. Ichikawa, H. Hayashi, and K. Nagasaka, "Neural network ensemble-based solar power generation short-term forecasting," *International Journal of Electrical and Computer Engineering*, vol. 3, no. 6, pp. 1258–1269, 2009, doi: 10.5281/zenodo.1070909.
- [24] J. Kajornrit, "A comparative study of optimization methods for improving artificial neural network performance," in *7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Oct. 2015, pp. 35–40, doi:

- 10.1109/ICITEED.2015.7408908.
- [25] S. Salisu, M. W. Mustafa, M. Mustapha, and O. O. Mohammed, "Solar radiation forecasting in Nigeria based on hybrid PSO-ANFIS and WT-ANFIS approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, pp. 3916–3926, Oct. 2019, doi: 10.11591/ijece.v9i5.pp3916-3926.
- [26] M. Farhadi and N. Mollayi, "Application of the least square support vector machine for point-to-point forecasting of the PV power," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 2205–2211, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2205-2211.
- [27] C. Voyant *et al.*, "Machine learning methods for solar radiation forecasting: a review," *Renewable Energy*, vol. 105, pp. 569–582, May 2017, doi: 10.1016/j.renene.2016.12.095.
- [28] C. Persson, P. Bacher, T. Shiga, and H. Madsen, "Multi-site solar power forecasting using gradient boosted regression trees," *Solar Energy*, vol. 150, pp. 423–436, Jul. 2017, doi: 10.1016/j.solener.2017.04.066.

## BIOGRAPHIES OF AUTHORS



**Alexandra Ilmarovna Khalyasmaa**    received the M.S. and Ph.D. degrees power industry from Ural Federal University named after the first President of Russia B.N. Yeltsin, Russia in 2009 and 2015, respectively. She is an associate professor at Electrical Engineering department, Ural Federal University named after the first President of Russia B.N. Yeltsin, Ekaterinburg, Russia and Power Plants department, Novosibirsk State Technical University, Novosibirsk, Russia. Her research lies in the sphere of machine learning application in power industry. She can be contacted at email: a.i.khalyasmaa@urfu.ru.



**Pavel Victorovich Matrenin**    received the M.S. and Ph.D. degrees information technologies from Novosibirsk State Technical University, Russia in 2014 and 2018, respectively. He is an Associate professor at Novosibirsk State Technical, Novosibirsk, Russia. His current research areas are stochastic optimization algorithms and machine learning in electric power systems. He can be contacted at email: matrenin.2012@corp.nstu.ru.



**Stanislav Andreevich Eroshenko**    received the M.S. and Ph.D. degrees power industry from Ural Federal University named after the first President of Russia B.N. Yeltsin, Russia in 2010 and from Novosibirsk State Technical University, Novosibirsk, Russia in 2020, respectively. He is a senior lecturer at Electrical Engineering Department, Ural Federal University named after the first President of Russia B.N. Yeltsin, Ekaterinburg, Russia and Power Plants department, Novosibirsk State Technical University, Novosibirsk, Russia. His research lies in the sphere of renewable energy systems forecasting. He can be contacted at email: s.a.eroshenko@urfu.ru.