

ANALYSIS AND MODELLING OF FAULTS IN SOLAR POWER PLANT

Emilija Bareikaitė, Aušrinė Budzevičiūtė, Rūta Levulienė
Institute of Applied Mathematics, Vilnius University, Vilnius, Lithuania
emilija.bareikaite@mif.stud.vu.lt, ausrine.budzeviciute@mif.stud.vu.lt, ruta.levulienė@mif.vu.lt

Abstract

Considering the rapid increase in electricity consumption and the necessity to transition to sustainable energy sources, the reliable operation of solar power plants has become particularly crucial. This study analyses time-series data from a privately operated active solar power plant, which is not publicly accessible, consisting of electrical parameters recorded every 5 minutes and corresponding fault labels. The primary focus is on the investigation of anomalies and the development of significant covariates that contribute to power plant failures. Subsequently, these variables are utilized in logistic regression models to predict operational disruptions, with the best model achieving an accuracy of 96.5%. The study is further enhanced by the application of a machine learning technique – random forests – which attained an accuracy of 98.4%. This research contributes to the more effective management of solar power plants, and aids in ensuring their reliable operation over the long term.

Keywords: machine learning, modelling, solar power plants, statistical analysis.

Analiza in modeliranje napak v sončnih elektrarnah

Izvleček

Ob upoštevanju hitrega naraščanja porabe električne energije ter potrebe po prehodu na trajnostne energetske vire je zanesljivo delovanje sončnih elektrarn postalo še posebej pomembno. V tej študiji so analizirani resnični podatki iz zasebne, javno nedostopne sončne elektrarne z namenom prepoznati dejavnike, ki vplivajo na delovanje sistema. Najprej je izvedena preliminarne analiza podatkov ter pregled metodologij, uporabljenih v dosednji literaturi. Glavni poudarek je na preučevanju anomalij in razvoju pomembnih spremenljivk, ki prispevajo k okvaram elektrarne. Te spremenljivke so nato uporabljene v modelih logistične regresije za napovedovanje motenj v delovanju, pri čemer je najboljši model dosegel natančnost 96,5 %. Raziskava je dodatno nadgrajena z uporabo metode strojnega učenja – naključnih gozdov – ki je dosegla natančnost 98,4 %. Ta raziskava prispeva k učinkovitejšemu upravljanju sončnih elektrarn ter pomaga zagotavljati njihovo zanesljivo delovanje na dolgi rok.

Keywords: strojno učenje, modeliranje, sončne elektrarne, statistična analiza.

1 INTRODUCTION

Electricity is essential in modern life, powering industry, transportation, and everyday activities. Global electricity consumption has been steadily increasing, reaching 27,000 TWh in 2023 [1]. Various sources contribute to electricity production, including coal, gas, oil, nuclear, hydropower, wind, solar, and other renewables. While coal and gas dominate electricity generation, they produce significant carbon dioxide emissions, prompting the rapid

growth of renewable energy. In 2023, renewables accounted for 30% of global electricity production [2], with solar power contributing 18% (1,630 TWh) [3].

Solar power plants are environmentally friendly, producing electricity without harmful emissions or noise. However, to maximize their contribution and reduce reliance on non-renewables, it is crucial to ensure their efficient operation and promptly detect faults. Monitoring and predicting failures not only optimize energy production but also reduce risks such as fires [4]. Despite their importance, fault detection can be challenging, as it requires high-resolution and precise observations to identify deviations from normal operation.

This study analyses real operational data from a solar power plant and develops predictive models to forecast alarms and identify the most significant factors affecting plant performance.

2 RELATED WORK

Several approaches have been proposed for fault detection in solar power plants. Early methods relied on statistical techniques such as hypothesis testing and control charts to identify deviations in energy production [5], [6], but they often cannot determine the exact cause or timing of a fault and typically ignore environmental influences.

Due to limited access to real-world data, many studies use simulated datasets to develop predictive models. Logistic regression [7], decision trees [8], and hybrid algorithms combining SVM and kNN [9] have been applied to simulated fault scenarios, achieving high accuracies. However, simulated data may not fully reflect real-world conditions, limiting model generalizability.

Studies using real solar power plant data have demonstrated higher practical applicability. Methods based on support vector machines (SVM), neural networks (BNN, CNN) [10], XGBoost [11], and combinations of random forests and kNN [12] achieved fault detection accuracies above 90%, confirming their effectiveness under operational conditions. In addition, image-based methods using RGB [13] and infrared (IRT) [14] data have been explored, although challenges such as limited data and long training times remain.

Unlike most prior studies that rely on simulations, this study uses real-world time-series data, including both environmental and electrical parameters, and applies classical statistical and machine learning models to enable accurate fault prediction. This provides a practical evaluation of model performance under actual operating conditions.

3 METHODOLOGY

3.1 Data Description and Preliminary Analysis

The analysis is based on private time-series data collected over a two-year period from a solar power plant consisting of eight inverters, each connected to twelve strings of solar modules and equipped with four MPPT (Maximum Power Point Tracking) controllers. The dataset includes electrical parameters such as voltage, current, and generated power, as well as a fault indicator recorded at 5-minute intervals. Preliminary exploration showed that the highest AC power output and the largest number of faults occurred during summer 2023, leading to a focused analysis on the three-month period from June to August. During this time, 923 fault warnings were recorded. Each fault warning has a different code, indicating any sort of deviation from normal operation, thus all warnings were assigned as one fault class. Since most warnings occurred during daytime hours with non-zero power generation, the analysis was restricted to daytime observations, which also reduced missing values and improved data quality. Exploratory analysis revealed differences in energy output among inverter strings and

in voltages across MPPT controllers. These differences were confirmed by Kruskal-Wallis tests ($p < 0.05$), indicating statistically significant variation between strings and controller voltages.

3.2 Data Preparation and Feature Engineering

Based on the initial analysis, literature review, and validated hypotheses, covariates were developed for model construction. The response variable was defined as a binary indicator of failure occurrence ("1" indicating a detected fault). Independent variables included meteorological and operational parameters. Hourly environmental data (irradiance, temperature, etc.) were aligned to the 5-minute interval of the electrical measurements. Operational features were derived from energy, voltage, and current readings, with short-term changes calculated over multiple intervals to capture dynamic system behaviour. Specifically:

- Energy changes were computed for each string over 5, 15, and 30-minute intervals.
- Voltage changes were calculated per inverter and MPPT controller over 5 and 15-minute intervals.
- Differences between the non-faulty 5th inverter and other inverters were computed using 5-minute voltage changes and categorized as "high" or "low" based on median thresholds.
- Original current values for each MPPT controller were retained.

Observations from consistently fault-free inverters were excluded. After these steps, the dataset comprised 92,389 observations and 60 variables, used for model development.

The dataset was highly imbalanced, with 913 failure cases versus 91,476 non-failure observations. To address this, the data were split into training (80%) and testing (20%) sets, with balancing applied only to the training data. A combination of undersampling the majority class and oversampling the minority class was used, including the generation of synthetic failure cases with added noise to prevent overfitting. Two balanced training datasets were created (smaller and larger) to assess model robustness under different class ratios (Table 1).

Table 1: Failure distribution before and after balancing.

Dataset Version	0 (No Failure)	1 (Failure)
Before balancing	73,180	731
After balancing (smaller training set)	10,000	2,031
After balancing (larger training set)	10,000	3,331

3.3 Predictive Modelling

Model development was conducted in parallel using both the smaller and larger training datasets. The initial logistic regression model included all 60 covariates. During training, warnings appeared indicating extreme predicted probabilities (close to 0 or 1), which can lead to unstable coefficients. Investigation showed that voltages from the first, second, and fourth MPPT controllers were responsible, and these variables were excluded from further analysis. Outlier analysis based on standardized residuals identified several observations outside the ± 3 range, which were removed. Cook's distance plots indicated no influential outliers (all < 1). Multicollinearity was then examined. Although principal component analysis (PCA) reduced collinearity, it also lowered predictive accuracy, so it was not adopted. Instead, correlated but

statistically significant variables were retained, and variance inflation factor (VIF) values were kept below 4 where possible. After resolving these issues, four logistic regression models were developed: two trained on the smaller dataset and two on the larger dataset (models 1–4). Model refinement involved removing statistically insignificant covariates ($p \geq 0.05$) and adding meaningful interaction terms, identified through both statistical testing and theoretical justification. The resulting models were then used for prediction and comparison.

In parallel, random forest models were developed using the same covariates and datasets. Several configurations were tested by varying the number of decision trees. Model parameters were selected to maximize sensitivity, as the primary objective was accurate detection of failure events. Other parameters, such as tree depth and minimum sample size, were left at default values after confirming that they had minimal impact on predictive stability.

3.4 Model Evaluation

Model evaluation was conducted using the testing datasets that reflected the original class distribution. Both logistic regression and random forest models were assessed using standard classification metrics, including accuracy, specificity, sensitivity, and F_2 scores, and the area under the ROC and PR curves (AUC). Classification thresholds for logistic regression models were tuned using a grid search between 0 and 1 to balance sensitivity and specificity. Since the prediction of rare failure events was the main focus of this study, sensitivity and the F_2 score were prioritized when comparing model performance, and models were ultimately selected based on their ability to maximize these metrics while maintaining high overall accuracy and stable AUC values.

4 RESULTS

Four logistic regression models were developed using training datasets with two different proportions of failure cases and varying covariate configurations. Table 2 summarizes their performance metrics, while Figure 1 presents the corresponding ROC and PR curves. Model 4 achieved the highest overall performance, with an accuracy of 96.5%, sensitivity of 70.9%, specificity of 96.8%, and an F_2 score of 0.45. Models 2 and 4 showed slightly higher accuracy and sensitivity values than Models 1 and 3. Since the larger training datasets had balanced classes, the probability threshold was unchanged from 0.5, which is a commonly used default decision threshold in logistic regression. For the smaller training datasets, the optimal probability threshold providing the best F_2 value without affecting specificity, which was 0.36, was determined through grid search. ROC and PR curves indicate similar overall performance among the four models, with minor differences in AUC values (Figure 1). Based on these results, Model 4 was selected as the final logistic regression model for further analysis due to its superior balance between sensitivity, specificity, and overall accuracy.

Table 2: Performance metrics of logistic regression models with different training data and thresholds.

Model	Training Data	Threshold	Accuracy	Sensitivity	Specificity	F_2 score
Model 1	Smaller	0.36	95.9%	67.6%	96.2%	0.40
Model 2	Smaller	0.36	96.4%	70.9%	96.7%	0.44
Model 3	Larger	0.5	96.2%	67.6%	96.5%	0.41
Model 4	Larger	0.5	96.5%	70.9%	96.8%	0.45

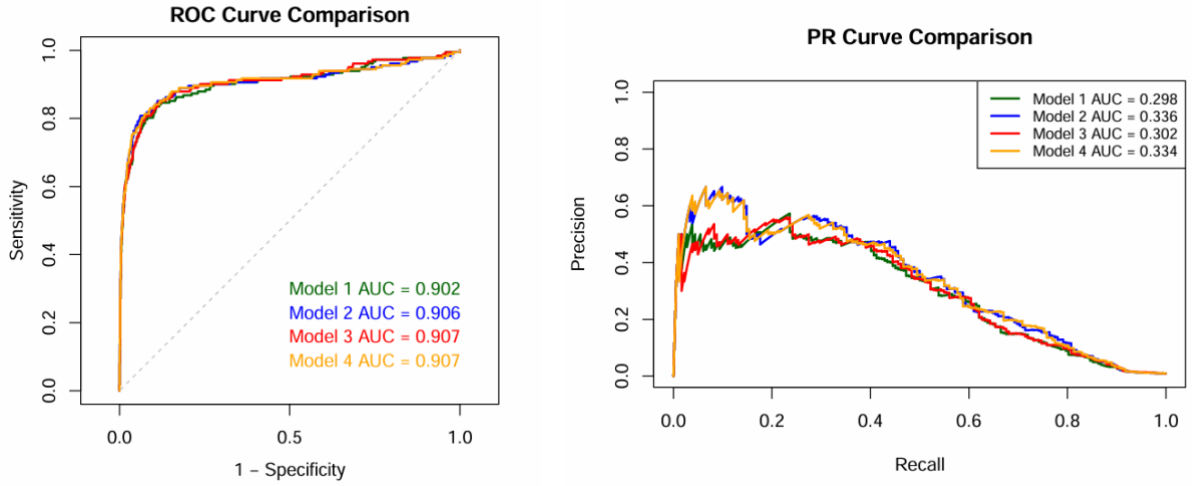


Figure 1: ROC and PR curves for logistic regression models.

Random forest models were trained using the same datasets and evaluated with different numbers of decision trees ranging from 40 to 250. The resulting performance metrics are presented in Table 3, and the ROC and PR curves are shown in Figure 2. For the smaller dataset, the highest sensitivity (78%) was achieved using 50 trees, with an accuracy of 98.4% and specificity of 98.6%. For the larger dataset, the best result was obtained with 150 trees, yielding a sensitivity of 77.5% and accuracy of 98.3%. In both cases, specificity exceeded 98%, and F_2 scores ranged between 0.61 and 0.63. Changes in the number of trees produced only minor differences in overall performance metrics. Among all tested configurations, the random forest model trained on the smaller dataset (50 trees) was identified as the final model, providing the highest sensitivity while maintaining high overall accuracy.

The most important variables of this model – those exerting the greatest influence on accuracy – are presented in Figure 3. Similar to the logistic regression models, meteorological variables play a significant role in the random forest model, as do the changes in energy of strings 2, 9, and 11, as well as the current of the third MPPT controller. However, most of the significant variables in the random forest model were not included in the aforementioned regression model (such as cloudiness, energy changes of other strings, voltage changes of other MPPT controllers, as well as the hour when the observation was recorded) due to high collinearity or lack of statistical significance.

Table 3: Performance metrics of random forest models with different training data and number of trees.

Number of Trees	Training Data	Accuracy	Sensitivity	Specificity	F_2 score
40	Smaller	98.3%	77.5%	98.5%	0.621
50	Smaller	98.4%	78%	98.6%	0.629
100	Smaller	98.4%	76.4%	98.6%	0.617
150	Smaller	98.4%	76.4%	98.6%	0.617
200	Smaller	98.4%	75.8%	98.6%	0.615
50	Larger	98.3%	76.4%	98.5%	0.613
100	Larger	98.3%	76.9%	98.5%	0.617
150	Larger	98.3%	77.5%	98.6%	0.622
200	Larger	98.4%	76.9%	98.6%	0.620
250	Larger	98.3%	76.4%	98.6%	0.615

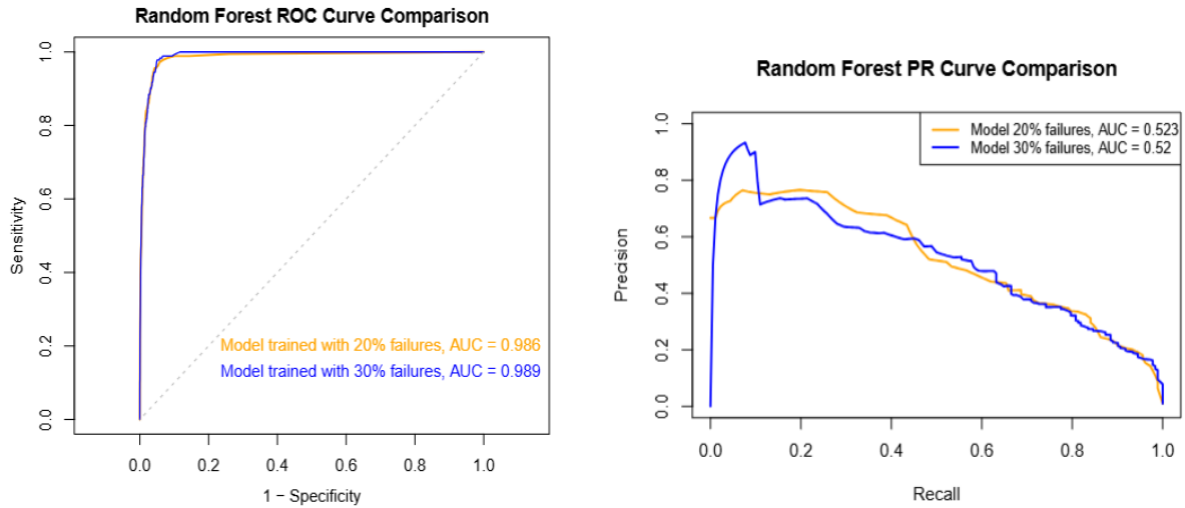


Figure 2: ROC and PR curves for random forest models.

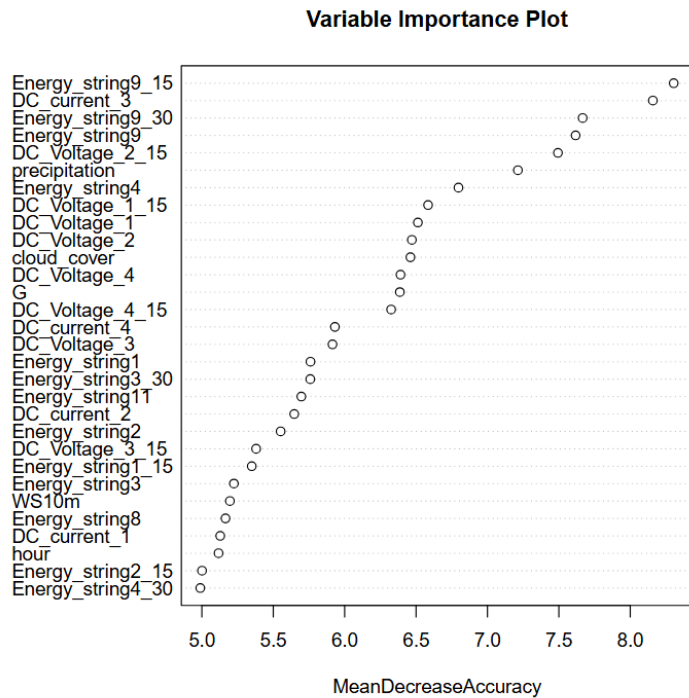


Figure 3: Variable importance plot of the random forest model with the highest sensitivity.

5 DISCUSSION

The findings of this study demonstrate that both statistical and machine learning models can effectively predict fault occurrences in photovoltaic systems when supported by thorough data preparation and feature engineering. Unlike many previous studies that rely on simulated datasets, this research was conducted using real operational data from an active solar power plant, providing a more realistic assessment of model performance under actual operating conditions. The random forest model outperformed logistic regression, confirming that nonlinear ensemble-based approaches are better suited for capturing complex relationships in photovoltaic system behaviour. This result aligns with prior research reporting the superior accuracy of machine learning techniques in solar fault detection. The inclusion of meteorological variables and time-dependent covariates further enhanced model

performance, emphasizing the importance of environmental conditions and temporal dynamics. While the models achieved strong predictive accuracy, certain limitations remain. The analysis was based on data from a single solar power plant, which may limit generalization to other installations operating under different environmental or technical conditions. Moreover, only two modelling approaches were investigated. Overall, these findings highlight the potential of data-driven methods for improving fault detection and predictive maintenance in renewable energy systems.

6 CONCLUSION

This study developed and compared logistic regression and random forest models for predicting faults in a photovoltaic power plant using two years of real operational data. Both models achieved high predictive accuracy: the logistic regression model reached 96.5% accuracy with 70.9% sensitivity, while the random forest model achieved 98.4% accuracy with 78% sensitivity, confirming its superior ability to capture nonlinear patterns and predict rare failure events.

A further examination of both statistical and machine learning approaches revealed that meteorological variables play a significant role in photovoltaic fault prediction, remaining statistically significant across both the logistic regression and random forest models. In addition, changes in the energy output of strings 2, 9, and 11 over different time intervals, derived during the data transformation phase, were identified as important predictors in both models, together with the current of MPPT controller 3. The variable importance analysis also highlighted differences between the modelling approaches. The logistic regression model emphasized interaction effects between energy output and irradiance, as well as interactions involving voltage and MPPT controller current, whereas the random forest model assigned greater importance to voltage fluctuations in MPPT controllers, controller current, observation time, and changes in the energy output of strings 1, 3, 4, and 8.

This study contributes to predictive maintenance in renewable energy by demonstrating that integrating real operational data with optimized feature engineering and model selection enhances the reliability and fault detection of photovoltaic systems. Future research could extend this work by incorporating data from multiple solar power plants, integrating additional predictors such as shading and temperature gradients, and applying advanced algorithms – such as gradient boosting or deep learning models – to further enhance predictive performance and generalizability.

References

- [1] Statista. (2023). *Global electricity consumption 1980-2023*. Retrieved from <https://www.statista.com/statistics/280704/world-power-consumption/>.
- [2] Ember. (2024). *World passes 30% renewable electricity milestone*. Retrieved from <https://ember-climate.org/press-releases/world-passes-30-renewable-electricity-milestone/>.
- [3] Ritchie, H., & Rosado, P. (2020). *Electricity Mix*. Our World in Data. Retrieved from <https://ourworldindata.org/electricity-mix>.
- [4] PVcase. (n.d.). *PV plant performance: challenges and solutions for large-scale solar projects*. Retrieved from <https://pvcase.com/blog/pv-performance#factors-influencing-pv-plant-performance>.
- [5] Vergura, S. (2018). *Hypothesis Tests-Based Analysis for Anomaly Detection in Photovoltaic Systems in the Absence of Environmental Parameters*. *Energies*, 11 (3), 485.

- [6] Garoudja, E., Harrou, F., Sun, Y., Kara, K., Chouder, A., & Silvestre, S. (2017). Statistical fault detection in photovoltaic systems. *Solar Energy*, 150, 485–499.
- [7] Ragul, S., Tamilselvi, S., Elamurugan, P., & Bharathidasan, S. (2024). Enhancing Photovoltaic System Resilience: A Logistic Regression Approach to Fault Diagnosis. In *Proceedings of the 2024 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*, pp. 1–5. IEEE.
- [8] Zhao, Y., Yang, L., Lehman, B., de Palma, J.-F., Mosesian, J., & Lyons, R. (2012). Decision tree-based fault detection and classification in solar photovoltaic arrays. In *Proceedings of the 2012 Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC)*, pp. 93–99. IEEE.
- [9] Rezgui, W., Mouss, L.-H., Mouss, N. K., Mouss, M. D., & Benbouzid, M. (2014). A smart algorithm for the diagnosis of short-circuit faults in a photovoltaic generator. In *Proceedings of the 2014 First International Conference on Green Energy (ICGE)*, pp. 139–143. IEEE.
- [10] Memon, S. A., Javed, Q., Kim, W.-G., Mahmood, Z., Khan, U., & Shahzad, M. (2022). A machine-learning-based robust classification method for PV panel faults. *Sensors*, 22(21), 8515.
- [11] Montes-Romero, J., Heinzle, N., Livera, A., Theocharides, S., Makrides, G., Sutterlueti, J., Ransome, S., & Georghiou, G. E. (2024). Novel data-driven health-state architecture for photovoltaic system failure diagnosis. *Solar Energy*, 279, 112820.
- [12] de Souza Silva, J. L., Mahmoudi, E., Carvalho, R. R. M., & dos Santos Barros, T. A. (2024). Classification of anomalies in photovoltaic systems using supervised machine learning techniques and real data. *Energy Reports*, 11, 4642–4656.
- [13] Espinosa, A. R., Bressan, M., & Giraldo, L. F. (2020). Failure signature classification in solar photovoltaic plants using RGB images and convolutional neural networks. *Renewable Energy*, 162, 249–256.
- [14] Ali, M. U., Khan, H. F., Masud, M., Kallu, K. D., & Zafar, A. (2020). A machine learning framework to identify the hotspot in photovoltaic module using infrared thermography. *Solar Energy*, 208, 643–651.

Emilija Bareikaitė, B.Sc. (Data Science). Emilija Bareikaitė graduated magna cum laude with a Bachelor's degree in Data Science from Vilnius University, Lithuania, where she will continue to pursue a Master's degree in Data Science. She is an ML/AI Researcher at AAI-Labs. Her research interests include natural language processing, machine learning, statistical modeling, and AI applications. She has authored and co-authored scientific publications and presented her research at national and international conferences.

Aušrinė Budzevičiūtė, B.Sc. (Data Science) received a Bachelor's degree in Data Science from the Faculty of Mathematics and Informatics at Vilnius University, Lithuania. Her research interests include machine learning, statistical modeling, and artificial intelligence. She has worked on the application of machine learning and statistical methods to solar power plant data analysis, including fault analysis and predictive modeling.

Rūta Levulienė, Ph.D. (Mathematics). Rūta Levulienė is an associate professor at the Faculty of Mathematics and Informatics of Vilnius University in Lithuania. Her areas of research interest are mathematical statistics and its applications. She has over 20 years of teaching experience in various statistics courses. Recently, she participated in the AI2SEP project, a scientific project related to the COVID-19 pandemic, where statistical and machine learning methods were applied to forecasting and modeling. She is a member of the Council of the Lithuanian Mathematical Society and the Lithuanian Statistical Society.