mathematics

MDPI

Editorial

Special Issue "Statistical Data Modeling and Machine Learning with Applications II"

Snezhana Gocheva-Ilieva *, Atanas Ivanov and Hristina Kulina

Department of Mathematical Analysis, University of Plovdiv Paisii Hilendarski, 4000 Plovdiv, Bulgaria; aivanov@uni-plovdiv.bg (A.I.); kulina@uni-plovdiv.bg (H.K.)
* Correspondence: snow@uni-plovdiv.bg

Currently, we are witnessing rapid progress and synergy between mathematics and computer science. This interaction leads to a greater effect in solving theoretical and applied problems. In this context, and following the good results of the first Special Issue, "Statistical Data Modeling and Machine Learning with Applications", a second edition covering the same 15 topics was announced at the end of 2021. The present Special Issue (SI), like the first, concerns the section "Mathematics and Computer Science". In total, 35 manuscripts were submitted for review. Of these, after a strict peer-review process by at least three anonymous reviewers, 15 articles were accepted and published.

Study [1] proposes effective models for forecasting the maximum hourly electricity consumption per day in Slovakia. Four types of models were built: gray models (GM(1,1)), nonlinear gray Bernoulli models (NGBM(1,1)), one ANN (based on a multi-layer feed-forward back-propagation (MLPFFBP) network), and a hybrid model. This approach includes a pre-processed data series that is used to build the transverse set of gray models, construct a special validation process for the MLPFFBP-ANN, and create a weighted hybrid model with GM(1,1) and the ANN. According to the three criteria, the models of the GM(1,1) set, ANN, and hybrid model reported better accuracy in forecasting values than officially provided forecasts, as the hybrid model has the best indicators.

In [2], a new simplified selective algorithm is proposed to increase the efficiency of ensemble methods based on decision trees and the index of agreement. This approach was demonstrated on real-world data to predict the 305-day milk yield of Holstein–Friesian cows. Using rotated principal components, classification and regression tree (CART) ensembles and bagging, and Arcing methods, a 30% reduction in the number of trees of the constructed selective ensembles was achieved. In addition, hybrid linear stacked models were built, yielding a 13.6% reduction in test set prediction errors compared to hybrid models with the nonselective ensembles.

The aim of paper [3] was to create an effective approach to detect and counter cyberattacks on Internet of Vehicular networks (IoV). An innovative, explainable neural network (xNN) model based on deep learning (DL), and in particular, Denial of Service (DoS) assaults, has been developed. To build the model, K-means was first applied for clustering, classification, and extraction of the best features for anomaly detection. After that, the xNN model was built to classify attacks. The model was tested on the two known empirical datasets, CICIDS2019 and UNSW-NB15. The calculated statistical indicators showed that the proposed feature-scoring approach outperforms the known published results in this field.

Publication [4] deals with single-index quantile regression (SIQR), a type of semi-parametric quantile regression for analyzing heterogeneous data. The quantile regression method with the SCAD penalty and Laplace error penalty were used to construct two sparse estimators for the considered SIQR. This leads to an efficient procedure for variable selection and parameter estimation. Theorems were proved for the N-consistency and oracle properties of the proposed estimators. Computer simulations with benchmark data

Citation: Gocheva-Ilieva, S.; Ivanov, A.; Kulina, H. Special Issue "Statistical Data Modeling and Machine Learning with Applications II". *Mathematics* **2023**, *11*, 2775. https://doi.org/10.3390/math11122775

Received: 15 June 2023
Accepted: 16 June 2023
Published: 20 June 2023

samples were performed. The method was shown to exhibit some resistance to heavy-tail errors and outliers while increasing the accuracy of parameter estimates.

Paper [5] presents a new computationally and highly efficient hybrid Bayesian network training algorithm called Forward with Early Dropping Hill Climbing (FEDHC), which is applicable to continuous or categorical variables. The algorithm applies the forward–backward-with-early-dropout (FBED) variable selection to each variable as a means of skeleton identification, followed by a hill-climb (HC) scoring phase. Another advantage of the proposed version of FEDHC is its robustness against outliers. FEDHC, PC Hill Climbing (PCHC), and Max–Min Hill Climbing (MMHC) were illustrated on two real cross-sectional datasets. A new, computationally efficient implementation of MMHC was also suggested.

In [6] the problem of estimating the graphs of conditional dependencies between variables under Gaussian settings is investigated. The authors present an improved Jewel 2.0 version of their previous Jewel 1.0 method. This was achieved on the basis of regression-based problem formulation with the appropriate minimization algorithm. Other contribution of the work is the proposed stability selection procedure that reduces the number of false positive scores in the estimated graphs. Simulation experiments were conducted.

The authors of [7] applied nonlinear autoregressive exogenous (NARX) networks coupled with an optimizing algorithm for wavelet filtering for modeling long-term dependencies and anomaly detection in noisy and nonstationary time series. A procedure using wavelet packets and stochastic thresholds was developed to approximate the decomposed components of the original data. The suggested wavelet filtering allows for the construction of a more accurate predictive NARX model. In addition, the NARX model was applied for anomaly detection. The results are demonstrated for ionospheric parameter time series prediction and ionospheric anomaly detection.

In paper [8], a method was developed for estimating the consolation prize of a slot machine jackpot using multidimensional integrals. Various modifications of the stochastic quasi-Monte Carlo approaches, such as lattice and digital sequences, Halton and Sobol sequences, and Latin hypercube sampling, were used to calculate the integrals. The expectations of the real consolation prize were evaluated, depending on the distribution of time and the number of players. The method was generalized for a multidimensional case. Computational experiments were performed.

Article [9] presents new theoretical and applied results in stochastic processes in spatial kinematics and line geometry for modeling some characteristics of 3D surfaces. The authors introduced theoretical principles on line-element geometry, kinematic surfaces, and the Gaussian process latent variable model (GPLVM). A method for surface approximation, unsupervised surface segmentation, and surface denoising in 3D modeling was described, which was based on the Bayesian GPLVM and the GPLVM with back constraints. The results were illustrated on sets with artificial and real-world objects.

In [10], a new method that aggregates five machine learning (ML) methods from different classification groups and a binary regression algorithm is proposed. The real-world task of predicting the impact of meteorological factors on the appearance of traffic accidents was solved. The most significant meteorological factors for road accidents were identified. The model was implemented as one of the agents in a two-agent system: agent 1 draws knowledge through ML from historically available data, and agent 2 deals with the same parameters, but in real-time. The suggested two-agent system can be implemented for providing early-warning alerts to citizens and traffic police, including through social media platforms.

The authors of [11] developed a novel, general multi-step-ahead strategy to forecast time series of air pollutants, extending the known multiple-input multiple-output (MIMO) strategy. The suggested strategy presupposes the availability of external independent forecasts for meteorological, atmospheric, and other variables, and continuously updated datasets. A new computational scheme was proposed for h-vector horizon prediction for each forward step. The strategy was applied to forecast the daily concentrations of

pollutants PM_{10}, SO_2, and NO_2 17 horizons ahead, with h = 10 days. Random forest (RF) and arcing (Arc-x4) ML algorithms were used for modeling. The comparison with the existing strategies showed the advantage of the proposed one.

Paper [12] presents a novel credit card fraud detection scheme, RaKShA, which is integrated with explainable artificial intelligence (XAI) and long short-term memory (LSTM), i.e., the X-LSTM model, and the output is verified via a smart contract (SC). The results are stored in the InterPlanetary File System (IPFS), which is referenced on the public blockchain network. The proposed approach addressed the limitation of traditional fraud detection by providing model interpretability, improved accuracy, security, and transparency. The X-LSTM model was found to increase the power of the LSTM model in detecting credit card financial fraud (FF) and to make the scheme scalable and adaptable, which helps users to protect themselves from FF.

Paper [13] presents an efficient one-stage model for automatic lung tumor detection in computed tomography (CT) images, called ELCT-YOLO. It was designed to solve the problem of scales and meet the requirements of real-time tumor detection. The ELCT-YOLO model implemented a specially designed neck structure and a novel Cascaded Refinement Scheme (CRS) to process context information. The results of empirical tests showing the advantages of the model were presented.

In [14] a Light Gradient Boosting Machine (LightGBM) model is utilized to classify and predict leisure time. The SHapley Additive exPlanation (SHAP) approach was applied to conduct feature importance analysis and influence mechanism analysis of factors from four perspectives: time allocation, demographics, occupation, and family characteristics. The results verified that the LightGBM model effectively predicts personal leisure time.

A two-layer autoencoder neural network architecture, singular-spectrum analysis (SSA) decomposition, and an adaptive anomaly detection algorithm (AADA) were used in [15] to process natural data of a complex, noisy nature. The AADA includes wavelet transforms whose accuracy is set with appropriate thresholds. These methods were applied for the analysis and detection of anomalous decreases that occurred before the geomagnetic disturbances. High-performance hybrid models were built for the study of cosmic ray data. The hybrid SSA-AADA models reach about 84% efficiency in anomaly detection, while the Autoencoder-AADA models reach about 87%.

To summarize, we should emphasize that the results of the published articles fully correspond to the formulated goal and topics of the SI, "Statistical Data Modeling and Machine Learning with Applications II". Their main contributions are classified in Table 1. We can note that the selected 15 topics are well covered. The hybrid models, machine learning algorithms, and nonparametric statistical modeling received the greatest interest.

Table 1. Classification by topics of the main contributions of the articles published in the SI.

Topic	Paper
Computational statistics	[4,6,10]
Dimensionality reduction and variable selection	[4–6]
Nonparametric statistical modeling	[4–6,8–10]
Supervised learning (classification, regression)	[10,12,13]
Clustering methods	[3,12]
Financial statistics and econometrics	[1,12]
Statistical algorithms	[4,8,9]
Time series analysis and forecasting	[1,7,10,11]
Machine learning algorithms	[2,5,7,10–13]
Decision trees	[2,14]
Ensemble methods	[2,10,11]
Neural networks	[1,3,5,7,12,15]
Deep learning	[3,7,12–14]
Hybrid models	[1,2,5,7,12–15]
Data analysis	[4,7,10,14,15]

In conclusion, new mathematical methods and approaches, new algorithms and research frameworks, and their applications aimed at solving diverse and nontrivial practical problems are proposed and developed in this SI. We believe that the chosen topics and results are attractive and useful for the international scientific community and will contribute to further research in the field of statistical data modeling and machine learning.

Funding: This research received no external funding.

Acknowledgments: The research activity of the first Guest Editor of this Special Issue was conducted within the framework of and was partially supported by the MES (Grant No. D01-168/28.07.2022) for NCDSC, as part of the Bulgarian National Roadmap on RIs.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pavlicko, M.; Vojteková, M.; Blažeková, O. Forecasting of Electrical Energy Consumption in Slovakia. *Mathematics* **2022**, *10*, 577. [CrossRef]
2. Gocheva-Ilieva, S.; Yordanova, A.; Kulina, H. Predicting the 305-Day Milk Yield of Holstein-Friesian Cows Depending on the Conformation Traits and Farm Using Simplified Selective Ensembles. *Mathematics* **2022**, *10*, 1254. [CrossRef]
3. Aziz, S.; Faiz, M.T.; Adeniyi, A.M.; Loo, K.-H.; Hasan, K.N.; Xu, L.; Irshad, M. Anomaly Detection in the Internet of Vehicular Networks Using Explainable Neural Networks (xNN). *Mathematics* **2022**, *10*, 1267. [CrossRef]
4. Song, Y.; Li, Z.; Fang, M. Robust Variable Selection Based on Penalized Composite Quantile Regression for High-Dimensional Single-Index Models. *Mathematics* **2022**, *10*, 2000. [CrossRef]
5. Tsagris, M. The FEDHC Bayesian Network Learning Algorithm. *Mathematics* **2022**, *10*, 2604. [CrossRef]
6. Angelini, C.; De Canditiis, D.; Plaksienko, A. Jewel 2.0: An Improved Joint Estimation Method for Multiple Gaussian Graphical Models. *Mathematics* **2022**, *10*, 3983. [CrossRef]
7. Mandrikova, O.; Polozov, Y.; Zhukova, N.; Shichkina, Y. Approximation and Analysis of Natural Data Based on NARX Neural Networks Involving Wavelet Filtering. *Mathematics* **2022**, *10*, 4345. [CrossRef]
8. Georgiev, S.; Todorov, V. Efficient Monte Carlo Methods for Multidimensional Modeling of Slot Machines Jackpot. *Mathematics* **2023**, *11*, 266. [CrossRef]
9. De Boi, I.; Ek, C.H.; Penne, R. Surface Approximation by Means of Gaussian Process Latent Variable Models and Line Element Geometry. *Mathematics* **2023**, *11*, 380. [CrossRef]
10. Aleksić, A.; Ranđelović, M.; Ranđelović, D. Using Machine Learning in Predicting the Impact of Meteorological Parameters on Traffic Incidents. *Mathematics* **2023**, *11*, 479. [CrossRef]
11. Gocheva-Ilieva, S.; Ivanov, A.; Kulina, H.; Stoimenova-Minova, M. Multi-Step Ahead Ex-Ante Forecasting of Air Pollutants Using Machine Learning. *Mathematics* **2023**, *11*, 1566. [CrossRef]
12. Raval, J.; Bhattacharya, P.; Jadav, N.K.; Tanwar, S.; Sharma, G.; Bokoro, P.N.; Elmorsy, M.; Tolba, A.; Raboaca, M.S. RaKShA: A Trusted Explainable LSTM Model to Classify Fraud Patterns on Credit Card Transactions. *Mathematics* **2023**, *11*, 1901. [CrossRef]
13. Ji, Z.; Zhao, J.; Liu, J.; Zeng, X.; Zhang, H.; Zhang, X.; Ganchev, I. ELCT-YOLO: An Efficient One-Stage Model for Automatic Lung Tumor Detection Based on CT Images. *Mathematics* **2023**, *11*, 2344. [CrossRef]
14. Wang, Q.; Jiang, Y. Leisure Time Prediction and Influencing Factors Analysis Based on LightGBM and SHAP. *Mathematics* **2023**, *11*, 2371. [CrossRef]
15. Mandrikova, O.; Mandrikova, B.; Esikov, O. Detection of Anomalies in Natural Complicated Data Structures Based on a Hybrid Approach. *Mathematics* **2023**, *11*, 2464. [CrossRef]