symmetry

MDPI

Editorial

Special Issue: Machine Learning and Data Analysis

Marcin Michalak [1,2]

[1] Department of Computer Networks and Systems, Silesian Univeristy of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland; marcin.michalak@polsl.pl
[2] Łukasiewicz Research Network—Institute of Innovative Technologies EMAG, ul. Leopolda 31, 40-189 Katowice, Poland; marcin.michalak@emag.lukasiewicz.gov.pl

This Special Issue contains 2 reviews and 17 research papers related to the following topics:

- Time series forecasting [1–5];
- Image analysis [6];
- Medical applications [7,8];
- Knowledge graph analysis [9,10];
- Cybersecurity [11–13];
- Traffic analysis [14,15];
- Agriculture [16];
- Environmental data analysis [17].

The authors of [1] focused on short time series forecasting in the domain of crime data (thefts, shoplifting, vehicular crimes, and burglaries in Mexico). The authors compared different combinations of model building (e.g., ARIMA, Simple Moving Averages, Artificial Neural Networks, and many other) with several approaches in predicting performance. As a result of the experiments carried out, two rankings using different prediction error measures (seasonal MAPE and the Friedman test) were compared. Both of them obtained the same rankings: SMA and ARIMA performed the best, FFORMA [18] was ranked in the middle, and the ANN and Holt–Winters models performed the worst. The presented results show promise as input for new heuristics to be applied on a larger set of time series.

In [2], a time series analysis was applied in a different manner: their prediction of the high stock dividend (HSD) was based on a sequence of typical machine learning approaches instead of state-of-the-art methods such as ARIMA or SMA. Four layers of genetic algorithms were used to find the optimal result. The first layer was responsible for a proper fitness function selection (based on random forests, XGBoost, and many others), while the second layer tried to find the optimal solution, the third layer determined the convergence range of the optimal result, and the fourth layer was responsible for dimensionality reduction. The most significant results of these experiments are the real dimensionality reduction and improvement in prediction evaluation measures.

An economical application of deep learning in time series prediction was presented in [3]. The authors focused their attention on predicting the daily prices of Bitcoin and gold. The newly proposed model combines neural networks with the basic economic indicators for better decision making.

In efforts to limit or stop climate change, renewable power sources have garnered more and more attention. Photovoltaic installations are surely one such renewable power source. However, the efficiency of such a source of clean energy depends strictly on the conditions of the outside environment, which makes it very hard to plan energy production. For that reason, three different levels of power production are being considered in this area: short term (from one hour to one week forward), middle term (up to one month), and long term (up to one year). The authors of [4] focused on short-term predictions.

The authors of [5] paid more attention to causal inferences in time series between dependent and independent variables. The authors built a new ensemble model and

Citation: Michalak, M. Special Issue: Machine Learning and Data Analysis. *Symmetry* **2023**, *15*, 1397. https://doi.org/10.3390/sym15071397

Received: 27 June 2023
Accepted: 30 June 2023
Published: 11 July 2023

compared its results with those of other tools, such as the Granger causality test, transfer entropy, and several more, on artificial multivariate data, containing linear and nonlinear dependent and independent variables.

In [6], the issue of applying the deep learning method to match the sea surface image to initial Beaufort scale levels was raised. The authors started with a broad review of existing benchmark datasets, as well as with a description of a various developed classification models. After conducting some experiments, they focused their attention on the GoogLeNet model—a convolutional neural network [19]. After some modification, increasing the classification accuracy, as well as decreasing the training time, became possible.

Deep learning methods were also used for predicting red blood cell (RBC) parameter values [7]. Better methods of RBC property modeling mean more reliable models of blood flow in general. The development of a new model became possible on the basis of the data simulated in the ESPREesSO [20] environment. Interestingly, the authors took both of the approaches—classification and regression—into consideration.

During the last three years, we have been focused on the COVID-19 pandemic. Hopefully, that threat has been more or less successfully dealt with in most of the world. However, we should also be paying more attention to leading causes of death in the world, particularly non-communicable chronic diseases [21]. In [8], a bio-inspired cuttlefish algorithm, CFA, and genetic algorithm, GA, were used to classify patients as suffering or not suffering from type 2 diabetes.

Knowledge extraction from graphs is a very important branch of data analysis, especially in domains such as recommendation systems, customer association analyses, or community detection. The work in [9] referred to a knowledge graph [22] (KG) analysis. The authors presented a new type of recommendation system called Personalized Relationships-Based Knowledge Graph for Recommender Systems with Dual-View Items. It utilizes a heterogeneous propagation strategy to gather information on higher-order user–item interactions and an attention mechanism to generate the weighted representation of entities. This paper also shows the results of the application of this approach to music, movie, and book recommendation scenarios, which showed increases compared with the results of state-of-the-art baselines.

Knowledge graphs are also the subject of analysis in [10]: a symmetric representation of KG. Joining the convolutional neural network with the TransE model, a tool called Joint Knowledge Representation Learning of Text Description and Knowledge Graph was obtained.

The authors of [11] focused on increasing the security of the Internet by detecting peer–to-peer botnets. Such a goal of data analysis is very popular in computer science, and many approaches have been developed based on a variety of machine learning and deep learning techniques, such as support vector machines, clustering techniques, Bayesian networks, decision trees, and many others. However, comparisons with multi-layer perceptron are lacking. The complete solution presented in the paper—called "PeerAmbush"—supports all steps (data construction, preparation, feature engineering, model building, and application) and provides satisfactory results on a benchmark dataset.

The authors of [12] also presented cybersecurity issues; however, they paid more attention to data preparation. The possibility of flow-based IDS feature enrichment was also raised for better detection of ICMPv6–DDoS attacks. The three main achievements in this paper are an increase in attack detection accuracy on enriched flow-based features, a reduction in feature sets required for such attack detections, and a wide comparison of suggested models applying a variety of machine learning techniques.

The authors of [13] partially addressed cybersecurity issues, as attempting to help detect frauds regarding credit cards. Unfortunately, for confidentiality issues, we cannot take a closer look at the input data model; however, the authors deal with highly imbalanced data (the percentage of fraud transaction does not exceed 0.2%). The application of a new method that combines an autoencoder with a light gradient boosting machine provided an observable increase in fraud detections.

Air traffic complexity was the point of interest in [14]. We have observed an increase in air traffic volume in recent decades. Even after a short reduction in air traffic volume during the COVID-19 pandemic, the amount of passengers and goods transported by means of air transport is still increasing year by year. In this paper, the authors attempt to answer the questions what is air traffic complexity and which air traffic data variables have greater impacts on increases in complexity? Machine learning techniques are used to find answers to these questions.

The authors of [15] focused more on the topological aspects of traffic analysis, proposing the Ollivier–Ricci curvature [23] to measure possible bottlenecks in the network.

The authors of [16] provided an interesting solution for increasing the accuracy of soil nutrient prediction by combining two tools: genetic algorithm (GA) and backpropagation neural networks (BPs). The authors compared their modified GA (IGA + BP) approach with a typical GA + BP approach, as well as with a BP-based prediction of the pH of soil, the total amount of nitrogen in soil, and the total amount of organic matter in soil, and the results are quite interesting.

In this Special Issue, we also have a paper that focuses on two aspects of data preprocessing [17]: imbalance of class distributions and missing values. The goal of these preprocessing aspects is to provide a good technique for assessing air quality (India Air Quality Index). Imbalance in the six classes varies from 5% (minority class) up to 36% (majority class), while the percentage of missing values for the seven variables reaches 26–29% for two of them. Two well-known techniques are involved in solving this imbalance: kNN and SMOTE.

Apart from the abovementioned research papers, this Special Issue contains two review papers. They are focused on informer methods used in a time series analysis [24] and the synergies between machine learning and neurorobotics [25]

Finally, I would like to congratulate all the authors of these papers on the acceptance of their work to this Special Issue, and I encourage the authors of rejected papers to improve their manuscripts and to then resubmit them to *Symmetry*.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Cruz-Nájera, M.A.; Treviño-Berrones, M.G.; Ponce-Flores, M.P.; Terán-Villanueva, J.D.; Castán-Rocha, J.A.; Ibarra-Martínez, S.; Santiago, A.; Laria-Menchaca, J. Short Time Series Forecasting: Recommended Methods and Techniques. *Symmetry* **2022**, *14*, 1231. [CrossRef]
2. Li, X.; Yu, Q.; Tang, C.; Lu, Z.; Yang, Y. Application of Feature Selection Based on Multilayer GA in Stock Prediction. *Symmetry* **2022**, *14*, 1415. [CrossRef]
3. Qi, Y.; Jiang, H.; Li, S.; Cao, J. ConvLSTM Coupled Economics Indicators Quantitative Trading Decision Model. *Symmetry* **2022**, *14*, 1896. [CrossRef]
4. Huang, Y.; Wu, Y. Short-Term Photovoltaic Power Forecasting Based on a Novel Autoformer Model. *Symmetry* **2023**, *15*, 238. [CrossRef]
5. Ma, Z.; Kemmerling, M.; Buschmann, D.; Enslin, C.; Lütticke, D.; Schmitt, R.H. A Data-Driven Two-Phase Multi-Split Causal Ensemble Model for Time Series. *Symmetry* **2023**, *15*, 982. [CrossRef]
6. Umair, M.; Hashmani, M.A.; Hussain Rizvi, S.S.; Taib, H.; Abdullah, M.N.; Memon, M.M. A Novel Deep Learning Model for Sea State Classification Using Visual-Range Sea Images. *Symmetry* **2022**, *14*, 1487. [CrossRef]
7. Molčan, S.; Smiešková, M.; Bachratý, H.; Bachratá, K. Computational Study of Methods for Determining the Elasticity of Red Blood Cells Using Machine Learning. *Symmetry* **2022**, *14*, 1732. [CrossRef]
8. Al-Tawil, M.; Mahafzah, B.A.; Al Tawil, A.; Aljarah, I. Bio-Inspired Machine Learning Approach to Type 2 Diabetes Detection. *Symmetry* **2023**, *15*, 764. [CrossRef]
9. Liu, Z.; Zhong, X.; Zhou, C. Personalized Relationships-Based Knowledge Graph for Recommender Systems with Dual-View Items. *Symmetry* **2022**, *14*, 2386. [CrossRef]
10. Xu, G.; Zhang, Q.; Yu, D.; Lu, S.; Lu, Y. JKRL: Joint Knowledge Representation Learning of Text Description and Knowledge Graph. *Symmetry* **2023**, *15*, 1056. [CrossRef]
11. Kabla, A.H.H.; Thamrin, A.H.; Anbar, M.; Manickam, S.; Karuppayah, S. PeerAmbush: Multi-Layer Perceptron to Detect Peer-to-Peer Botnet. *Symmetry* **2022**, *14*, 2483. [CrossRef]
12. Elejla, O.E.; Anbar, M.; Hamouda, S.; Belaton, B.; Al-Amiedy, T.A.; Hasbullah, I.H. Flow-Based IDS Features Enrichment for ICMPv6-DDoS Attacks Detection. *Symmetry* **2022**, *14*, 2556. [CrossRef]

13. Du, H.; Lv, L.; Guo, A.; Wang, H. AutoEncoder and LightGBM for Credit Card Fraud Detection Problems. *Symmetry* **2023**, *15*, 870. [CrossRef]
14. Pérez Moreno, F.; Gómez Comendador, V.F.; Delgado-Aguilera Jurado, R.; Zamarreño Suárez, M.; Janisch, D.; Arnaldo Valdés, R.M. Determination of Air Traffic Complexity Most Influential Parameters Based on Machine Learning Models. *Symmetry* **2022**, *14*, 2629. [CrossRef]
15. Han, X.; Zhu, G.; Zhao, L.; Du, R.; Wang, Y.; Chen, Z.; Liu, Y.; He, S. Ollivier–Ricci Curvature Based Spatio-Temporal Graph Neural Networks for Traffic Flow Forecasting. *Symmetry* **2023**, *15*, 995. [CrossRef]
16. Liu, Y.; Jiang, C.; Lu, C.; Wang, Z.; Che, W. Increasing the Accuracy of Soil Nutrient Prediction by Improving Genetic Algorithm Backpropagation Neural Networks. *Symmetry* **2023**, *15*, 151. [CrossRef]
17. Chandra, W.; Suprihatin, B.; Resti, Y. Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction. *Symmetry* **2023**, *15*, 887. [CrossRef]
18. Montero-Manso, P.; Athanasopoulos, G.; Hyndman, R.J.; Talagala, T.S. FFORMA: Feature-based forecast model averaging. *Int. J. Forecast.* **2020**, *36*, 86–92. [CrossRef]
19. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
20. Arnold, A.; Lenz, O.; Kesselheim, S.; Weeber, R.; Fahrenberger, F.; Roehm, D.; Košovan, P.; Holm, C. ESPResSo 3.1: Molecular Dynamics Software for Coarse-Grained Models. In *Proceedings of the Meshfree Methods for Partial Differential*; Equations, V.I., Griebel, M., Schweitzer, M.A., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 1–23.
21. Yach, D.; Hawkes, C.; Gould, C.L.; Hofman, K.J. The Global Burden of Chronic DiseasesOvercoming Impediments to Prevention and Control. *JAMA* **2004**, *291*, 2616–2622. [CrossRef] [PubMed]
22. Ehrlinger, L.; Wöß, W. Towards a Definition of Knowledge Graphs. *Semantics* **2016**, *48*, 2.
23. Ollivier, Y. Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.* **2009**, *256*, 810–864. [CrossRef]
24. Zhu, Q.; Han, J.; Chai, K.; Zhao, C. Time Series Analysis Based on Informer Algorithms: A Survey. *Symmetry* **2023**, *15*, 951. [CrossRef]
25. Lin, C.L.; Zhu, Y.H.; Cai, W.H.; Su, Y.S. Recent Synergies of Machine Learning and Neurorobotics: A Bibliometric and Visualized Analysis. *Symmetry* **2022**, *14*, 2264. [CrossRef]