technologies

Editorial

Editorial for the Special Issue "Data Science and Big Data in Biology, Physical Science and Engineering"

Mohammed Mahmoud

Department of Computer Science, University of Jamestown, Jamestown, ND 58405, USA; prof.mahmoud@uj.edu

Big Data analysis is one of the most contemporary areas of development and research in the present day. Tremendous amounts of data are generated every single day from digital technologies and modern information systems, such as cloud computing and Internet of Things (IoT) devices. Analysis of these enormous amounts of data became a essential and requires a large amount of effort in order to extract valuable knowledge for decision making which, in turn, will help in both academia and industry.

Big Data and Data Science have appeared due to the significant need for generating, storing, organizing and processing immense amounts of data. Data Scientists strive to use Artificial Intelligence (AI) and Machine Learning (ML) approaches and models to enable computers to detect and identify what the data represent and to detect patterns more quickly, efficiently and reliably than humans.

The goal behind this Special Issue is to explore and discuss various principles, tools and models in the context of Data Science, in addition to diverse and varied concepts and techniques regarding Big Data in Biology, Chemistry, Biomedical Engineering, Physics, Mathematics and other areas.

In this Special Issue, we present 12 papers that span a wide range of topics relating to Data Science, Big Data, machine learning, deep learning, Artificial Intelligence (AI) and cybersecurity.

In [1], the authors explored the application of various machine learning techniques for predicting customer churn in the telecommunications sector. They utilized a publicly accessible dataset and implemented several models, including Artificial Neural Networks, Decision Trees, Support Vector Machines, Random Forests, Logistic Regression and gradient boosting techniques (XGBoost, LightGBM and CatBoost). To mitigate the challenges posed by imbalanced datasets, the authors adopted different data sampling strategies, namely, SMOTE, SMOTE combined with Tomek Links and SMOTE combined with Edited Nearest Neighbors. Moreover, hyperparameter tuning was employed to enhance model performance. Their evaluation employed standard metrics, such as Precision, Recall, F1-score and the Receiver Operating Characteristic Area Under Curve (ROC AUC). In terms of the F1-score metric, CatBoost demonstrates superior performance compared to other machine learning models, achieving an outstanding 93% following the application of Optuna hyperparameter optimization. In the context of the ROC AUC metric, both XGBoost and CatBoost exhibit exceptional performance, recording remarkable scores of 91%. This achievement for XGBoost is attained after implementing a combination of SMOTE with Tomek Links, while CatBoost reaches this level of performance after the application of Optuna hyperparameter optimization.

In [2], the authors discussed the "Get Real Get Better" (GRGB) approach to implementing agile program management in the U.S. Navy, supported by advanced data analytics and Artificial Intelligence (AI). GRGB was designed as a set of foundational principles to advance Navy culture and support its core values. This article identifies a need for a more informed and efficient approach to program management by highlighting the benefits of implementing comprehensive data analytics that leverage recent advances in cloud computing and machine learning. The Jupiter enclave within Advana implemented by the

Citation: Mahmoud, M. Editorial for the Special Issue "Data Science and Big Data in Biology, Physical Science and Engineering". *Technologies* **2024**, *12*, 8. https://doi.org/10.3390/technologies12010008

Received: 18 December 2023
Revised: 27 December 2023
Accepted: 4 January 2024
Published: 8 January 2024

U.S. Navy is also discussed. The presented approach represents a practical framework that cultivates a "Get Real Get Better" mindset for implementing agile program management in the U.S. Navy.

As the manufacturing industry advances towards Industry 5.0, which heavily integrates advanced technologies such as cyber-physical systems, Artificial Intelligence, and the Internet of Things (IoT), the potential for web-based attacks increases. Cybersecurity concerns remain a crucial challenge for Industry 5.0 environments, where cyber attacks can cause devastating consequences, including production downtime, data breaches and even physical harm. To address this challenge, the third paper [3] proposed an innovative deep-learning methodology for detecting web-based attacks in Industry 5.0. Convolutional neural networks (CNNs), recurrent neural networks (RNNs) and transformer models are examples of deep learning techniques that are investigated in this study for their potential to effectively classify attacks and identify anomalous behavior. The proposed transformer-based system outperforms traditional machine learning methods and existing deep learning approaches in terms of accuracy, precision and recall, demonstrating the effectiveness of deep learning for intrusion detection in Industry 5.0. The study's findings highlighted the superiority of the proposed transformer-based system, outperforming previous approaches in accuracy, precision and recall. This highlights the significant contribution of deep learning in addressing cybersecurity challenges in Industry 5.0 environments, ensuring the protection of critical infrastructure and sensitive data.

Self-directed learning and self-design became unexpectedly popular and common during the COVID-19 era. Learners are encouraged to take charge of their learning and, often, the opportunity to independently design their learning experience. The fourth paper [4] illustrates the use of technology in teaching and learning technology with a central theme of promoting self-directed learning with engaging self-design for both educators and learners. The technology used includes existing tools such as web page design, Learning Management Systems (LMS), project management tools and basic programming foundations and concepts of Big Data and databases. In addition, end-users and developers can create their own tools with simple coding. Planning techniques, such as Visual Plan Construct Language with its embedded AI, are used to integrate course material and rubrics with time management. Educators may use project management tools instead. The research proposes a self-directed paradigm with self-designed resources using the existing technology with LMS modules, discussions and self-tests. The research establishes its criteria for ensuring the quality of content and design, known as 7x2C. Additionally, other criteria for analysis, such as Design Thinking, are included. The approach is examined for a technology-based business course in creating an experiential learning system for COVID-19 awareness. Likewise, among other projects, an environment for educating learners about diabetes and obesity has been designed. The project is known as Sunchoke, which has a theme of Grow, Eat and Heal. Educators can use their own content and rubrics to adapt this approach to their own customized teaching methods.

Deep neural networks (DNNs), the integration of neural networks (NNs) and deep learning (DL), have proven highly efficient in executing numerous complex tasks, such as data and image classification. Because the multilayer in a nonlinearly separable data structure is not transparent, it is critical to develop a specific data classification model from a new and unexpected dataset. In the fifth paper [5], the authors proposed a novel approach using the concepts of DNNs and decision trees (DTs) for classifying nonlinear data. They first developed a decision tree-based neural network (DTBNN) model. Next, they extended their model to a decision tree-based deep neural network (DTBDNN), in which the multiple hidden layers in a DNN are utilized. By using a DNN, the DTBDNN model achieved higher accuracy compared to the related and relevant approaches. Their proposal achieved the optimal trainable weights and bias to build an efficient model for nonlinear data classification by combining the benefits of DTs and NNs. By conducting in-depth performance evaluations, they demonstrated the effectiveness and feasibility of the proposal by achieving good accuracy over different datasets.

Organizations must quickly adapt their processes to understand the dynamic nature of modern business environments. As highlighted in the literature [6], centralized governance supports decision making and performance measurement processes in technology companies. For this reason, a reliable decision-making system with an integrated data model that enables the rapid collection and transformation of data stored in heterogeneous and different sources is needed. Therefore, the sixth paper [6] proposed the design of a data model to implement data-driven governance through a literature review of adopted approaches. The lack of a standardized procedure and a disconnection between theoretical frameworks and practical application has emerged. This paper documented the suggested approach following these steps: (i) mapping of monitoring requirements to the data structure, (ii) documentation of ER diagram design, and (iii) reporting dashboards used for monitoring and reporting. The paper helped fill the gaps highlighted in the literature by supporting the design and development of a DWH data model coupled with a BI system. The application prototype shows benefits for top management, particularly those responsible for governance and operations, especially risk monitoring, audit compliance, communication, knowledge sharing on strategic areas of the company, and identification and implementation of performance improvements and optimizations.

The discretization of continuous attributes in a dataset is an essential step before the Rough-Set-Theory (RST)-based classification process is applied. There are many methods for discretization, but few have linked RST instruments from the beginning of the discretization process. The objective of the seventh paper [7] was to propose a method to improve the accuracy and reliability of the RST-based classifier model by involving RST instruments at the beginning of the discretization process. In the proposed method, a k-means-based discretization method optimized with a genetic algorithm (GA) was introduced. Four datasets taken from UCI were selected to test the performance of the proposed method. The evaluation of the proposed discretization technique for RST-based classification was performed by comparing it to other discretization methods, i.e., equal frequency and entropy-based. The performance comparison among these methods was measured by the number of bins and rules generated and by their accuracy, precision and recall. A Friedman test, continued with post hoc analysis, was also applied to measure the significance of the difference in performance. The experimental results indicate that, in general, the performance of the proposed discretization method is significantly better than the other compared methods.

The eighth paper [8] presented a novel machine learning approach to predict sex in bioarchaeological records. Eighteen cranial interlandmark distances and five maxillary dental metric distances were recorded from n = 420 human skeletons from the necropolises at Alfedena (600–400 BCE) and Campovalano (750–200 BCE and 9–11th Centuries CE) in central Italy. A generalized low rank model (GLRM) was used to impute missing data and the Receiver Operating Characteristic Area Under Curve (AUC-ROC) with 20-fold stratified cross-validation was used to evaluate the predictive performance of eight machine learning algorithms on different subsets of the data. Additional perspectives such as this one show strong potential for sex prediction in bioarchaeological and forensic anthropological contexts. Furthermore, GLRMs have the potential to handle missing data in ways previously unexplored in the discipline. Although the results of this study look promising (highest AUC-ROC = 0.9722 for predicting binary male/female sex), the main limitation is that the sexes of the individuals included were not known but were estimated using standard macroscopic bioarchaeological methods. However, future research should apply this machine learning approach to known-sex reference samples in order to better understand its value, along with the more general contributions that machine learning can make to the reconstruction of past human lives.

Deep learning has been the answer to many machine learning problems during the past two decades. However, it comes with two significant constraints: dependency on extensive labeled data and training costs. Transfer learning in deep learning, known as Deep Transfer Learning (DTL), attempts to reduce such reliance and costs by reusing obtained

knowledge from source data/task in training on a target data/task. Most applied DTL techniques are network/model-based approaches. These methods reduce the dependency of deep learning models on extensive training data and drastically decrease training costs. Moreover, the training cost reduction makes DTL viable on edge devices with limited resources. Like any new advancement, DTL methods have their own limitations, and a successful transfer depends on specific adjustments and strategies for different scenarios. The ninth paper [9] reviewed the concept, definition, and taxonomy of deep transfer learning and well-known methods. It investigated DTL approaches by reviewing applied DTL techniques of the past five years and a couple of experimental analyses of DTLs to discover the best practice for using DTL in different scenarios. Moreover, the limitations of DTLs (catastrophic forgetting dilemma and overly biased pre-trained models) were discussed, along with possible solutions and research trends.

Despite best efforts, the loss of biodiversity has continued at a pace that constitutes a major threat to the efficient functioning of ecosystems. Curbing the loss of biodiversity and assessing its local and global trends requires a vast amount of datasets from a variety of sources. Although the means for generating, aggregating and analyzing big datasets to inform policies are now within reach of the scientific community, the data-driven nature of a complex multidisciplinary field such as biodiversity science necessitates an overarching framework for engagement. In the tenth paper [10], the authors proposed such a schematic based on the life cycle of data to interrogate the science. The framework considers data generation and collection, storage and curation, access and analysis and, finally, communication as distinct yet interdependent themes for engaging biodiversity science for the purpose of making evidenced-based decisions. The authors summarized historical developments in each theme, including the challenges and prospects, and offered some recommendations based on best practices.

The warehousing industry is faced with increasing customer demands and growing global competition. A major factor in the efficient operation of warehouses is the strategic storage location assignment of arriving goods, termed the dynamic storage location assignment problem (DSLAP). The eleventh paper [11] presented a real-world case of the DSLAP, in which deep reinforcement learning (DRL) is used to derive a suitable storage location assignment strategy to decrease transportation costs within the warehouse. The DRL agent is trained on historic data of storage and retrieval operations gathered over one year of operation. An evaluation of the agent using new data of the past two months showed a 6.3% decrease in incurring costs compared to the currently utilized storage location assignment strategy, which is based on manual ABC classifications. Hence, DRL proves to be a competitive solution for the DSLAP and related problems in the warehousing industry.

The software selection process in the context of a big company is not an easy task. In the business intelligence area, this decision is critical since the resources needed to implement the tool are huge and necessitate the participation of all organization actors. In the twelfth paper [12], the authors proposed to adopt the systemic quality model to perform a neutral comparison between four business intelligence self-service tools. To assess the quality, they considered eight characteristics and eighty-two metrics. They built a methodology to evaluate self-service BI tools, adapting the systemic quality model. As an example, they evaluated four tools that were selected from all business intelligence platforms, following a rigorous methodology. Through the assessment, they obtained two tools with the maximum quality level. To acquire the differences between them, they were more restrictive increasing the level of satisfaction. Finally, they obtained a unique tool with the maximum quality level, while the other one was rejected according to the rules established in the methodology. The methodology works well for this type of software, helping in the detailed analysis and neutral selection of the final software to be used for the implementation.

Acknowledgments: I would like to take the opportunity to thank all the authors for submitting their work and contributing to the journal, as well as their passion for research. I would also like to extend a special thank you to the reviewers for their dedication in reading the submitted papers and providing useful comments that helped support their entry into the Special Issue. It was an absolute pleasure reviewing the submitted work for the Special Issue of the *Technologies* journal.

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

References

1. Imani, M.; Arabnia, H.R. Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis. *Technologies* **2023**, *11*, 167. [CrossRef]
2. Haase, J.; Walker, P.B.; Berardi, O.; Karwowski, W. Get Real Get Better: A Framework for Developing Agile Program Management in the U.S. Navy Supported by the Application of Advanced Data Analytics and AI. *Technologies* **2023**, *11*, 165. [CrossRef]
3. Salam, A.; Ullah, F.; Amin, F.; Abrar, M. Deep Learning Techniques for Web-Based Attack Detection in Industry 5.0: A Novel Approach. *Technologies* **2023**, *11*, 107. [CrossRef]
4. Ebrahimi, A. Self-Directed and Self-Designed Learning: Integrating Imperative Topics in the Case of COVID-19. *Technologies* **2023**, *11*, 85. [CrossRef]
5. Arifuzzaman, M.; Hasan, M.R.; Toma, T.J.; Hassan, S.B.; Paul, A.K. An Advanced Decision Tree-Based Deep Neural Network in Nonlinear Data Classification. *Technologies* **2023**, *11*, 24. [CrossRef]
6. Biagi, V.; Russo, A. Data Model Design to Support Data-Driven IT Governance Implementation. *Technologies* **2022**, *10*, 106. [CrossRef]
7. Dwiputranto, T.H.; Setiawan, N.A.; Adji, T.B. Rough-Set-Theory-Based Classification with Optimized *k*-Means Discretization. *Technologies* **2022**, *10*, 51. [CrossRef]
8. Muzzall, E. A Novel Ensemble Machine Learning Approach for Bioarchaeological Sex Prediction. *Technologies* **2021**, *9*, 23. [CrossRef]
9. Iman, M.; Arabnia, H.R.; Rasheed, K. A Review of Deep Transfer Learning and Recent Advancements. *Technologies* **2023**, *11*, 40. [CrossRef]
10. Musvuugwa, T.; Dlomu, M.G.; Adebowale, A. Big Data in Biodiversity Science: A Framework for Engagement. *Technologies* **2021**, *9*, 60. [CrossRef]
11. Waubert de Puiseau, C.; Nanfack, D.T.; Tercan, H.; Löbbert-Plattfaut, J.; Meisen, T. Dynamic Storage Location Assignment in Warehouses Using Deep Reinforcement Learning. *Technologies* **2022**, *10*, 129. [CrossRef]
12. Orcajo Hernández, J.; Fonseca i Casas, P. Business Intelligence's Self-Service Tools Evaluation. *Technologies* **2022**, *10*, 92. [CrossRef]