ARTICLE

DOI 10.36074/grail-of-science.08.12.2023.33

XGBOOST IN ENVIRONMENTAL ECOLOGY: A POWERFUL TOOL FOR SUSTAINABLE INSIGHTS

SCIENTIFIC RESEARCH GROUP:

Tymoteusz Miller
PhD in biological sciences, assistant Professor at institute of Marine and Environmental Sciences
University of Szczecin. Polish Society of Bioinformatics and Data Science BIODATA, Szczecin, Poland

Polina Kozlovska
Bachelor of Genetics and Experimental Biology, Faculty of Physical, Mathematical and Natural Sciences
University of Szczecin. Polish Society of Bioinformatics and Data Science BIODATA, Szczecin, Poland

Adrianna Krzemińska
Bachelor of Genetics and Experimental Biology, Faculty of Physical, Mathematical and Natural Sciences
University of Szczecin. Polish Society of Bioinformatics and Data Science BIODATA, Szczecin, Poland

Klaudia Lewita
Bachelor of Genetics and Experimental Biology,
Faculty of Physical, Mathematical and Natural Sciences
University of Szczecin. Polish Society of Bioinformatics and Data Science BIODATA, Szczecin, Poland

Julia Biedrzycka
1st year student of Oceanography,
Faculty of Physical, Mathematical and Natural Sciences
University of Szczecin, Poland

Karolina Geroch
1st year student of Oceanography,
Faculty of Physical, Mathematical and Natural Sciences
University of Szczecin, Poland

Summary. *Environmental ecology stands at the forefront of understanding and addressing the challenges posed by a rapidly changing world. In this context, machine learning, particularly the XGBoost algorithm, has emerged as a pivotal tool, offering unparalleled accuracy and adaptability. This article delves into the origins and workings of XGBoost, highlighting its applications in predicting species distributions, assessing habitat suitability, and modeling climate change impacts. While the benefits of XGBoost, such as high predictive power and robustness to noisy data, are emphasized, the article also sheds light on potential challenges like overfitting and interpretability. The conclusion underscores the importance of a holistic approach, combining domain knowledge with algorithmic prowess, to harness the full potential of XGBoost in environmental ecology.*

Keywords: *XGBoost, Environmental Ecology, Machine Learning, Species Distribution, Habitat Suitability, Climate Change Modeling, Predictive Analysis, Data-driven Conservation*

1. Introduction

Machine learning, a subset of artificial intelligence, has revolutionized numerous fields with its ability to 'learn' from data and make predictions or decisions without being explicitly programmed. One such field that has greatly benefited from the advancements in machine learning is environmental ecology. Through the analysis of vast and complex datasets, machine learning algorithms can assist researchers in understanding patterns, predicting future trends, and making informed decisions to protect our environment [1,2].

Among the myriad of machine learning algorithms, XGBoost stands out due to its efficiency and accuracy. Originally designed for speed and performance, XGBoost has become a favorite among data scientists and researchers alike. Its adaptability and robustness make it particularly signifi ant in the realm of environmental ecology, where data can often be noisy or incomplete [3,4].

2. What is XGBoost?

In the vast landscape of machine learning algorithms, XGBoost, or Extreme Gradient Boosting, has emerged as a frontrunner. But what exactly is XGBoost, and why has it gained such prominence?

Historical Background and Development:

XGBoost was introduced as an improvement over the traditional Gradient Boosting Machines (GBM). Developed by Tianqi Chen, it was initially designed to optimize large-scale tree boosting. The algorithm quickly gained popularity after winning several Kaggle competitions, showcasing its prowess in handling diverse datasets [5,6].

Key Features and Advantages:

- Scalability: XGBoost can efficiently handle large datasets, making it suitable for real-world applications where data is abundant.
- Parallel Processing: Unlike traditional GBM, XGBoost utilizes parallel processing, speeding up the training process [3].
- Regularization: Built-in L1 (Lasso Regression) and L2 (Ridge Regression) regularization helps in preventing overfitt ng [7].
- Flexibility: It can handle regression, classification, ranking, and user-defined prediction tasks [8].

Basic Working Principle:

At its core, XGBoost is an ensemble learning method. It builds multiple decision trees and combines them to produce a final prediction. The "boosting" in its name refers to the algorithm's ability to convert weak learners into strong learners by iteratively correcting the errors of the previous trees [9,10].

3. Applications of XGBoost in Environmental Ecology

The versatility of XGBoost has made it a valuable tool in the domain of environmental ecology. Here are some notable applications:

Predicting Species Distribution:

Understanding where species live and how they interact with their environment is crucial for conservation effo ts. XGBoost can analyze various environmental factors, such as temperature, precipitation, and land use, to predict the distribution of species across diff rent habitats [11,12] (Fig. 1).

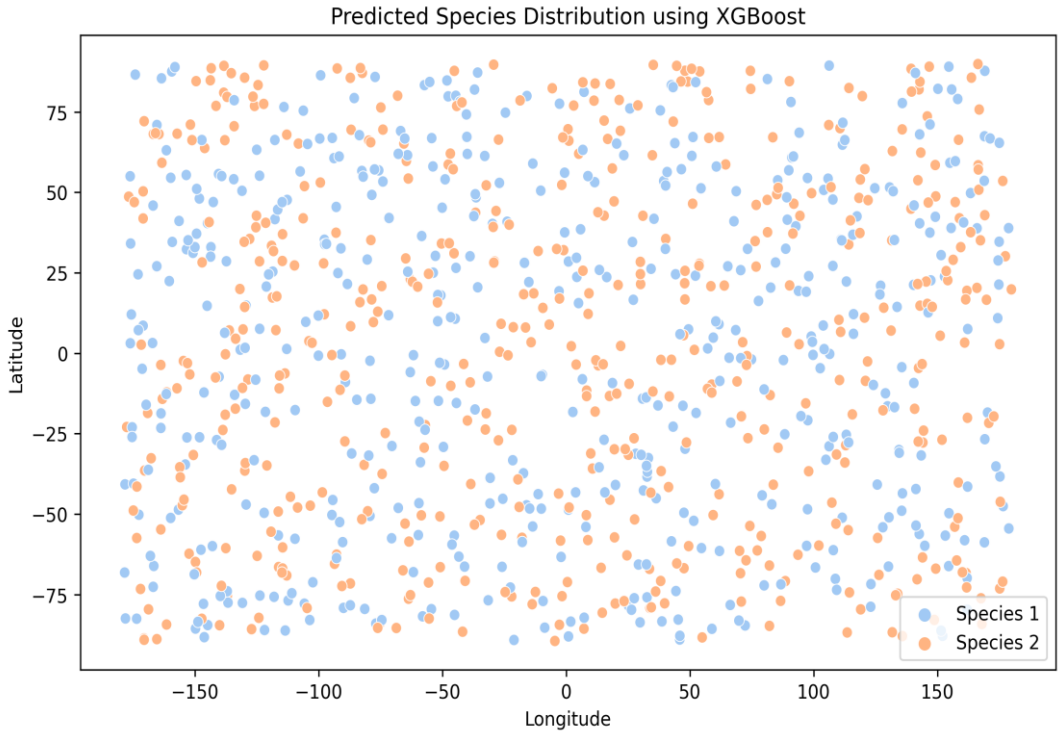

Fig. 1: Example Species Distribution Prediction using XGBoost

Habitat Suitability Modeling:

Determining the suitability of a habitat for a particular species is essential for reintroduction programs and conservation planning. By analyzing factors like vegetation, soil type, and human disturbances, XGBoost can provide insights into which areas are most suitable for a given species [12,13] (Fig. 2).

Assessing the Impact of Climate Change on Biodiversity:

Climate change poses a signifi ant threat to biodiversity. XGBoost can be used to model how changes in climate variables, such as temperature and rainfall, might aff ct the distribution and abundance of various species [14,15] (Fig. 3).

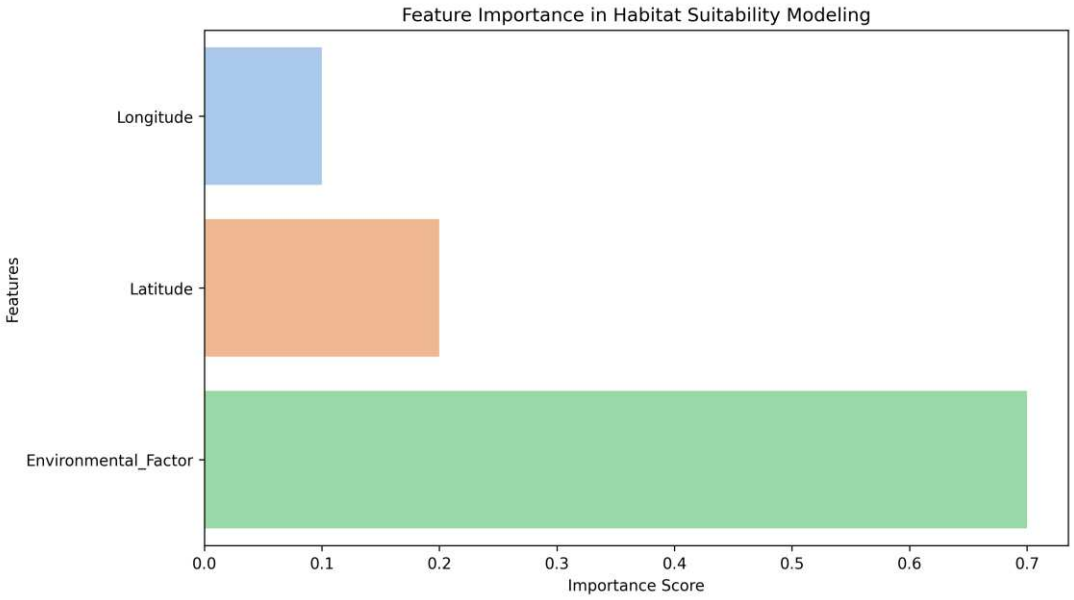

Fig. 2: Example Feature Importance in Habitat Suitability Modeling

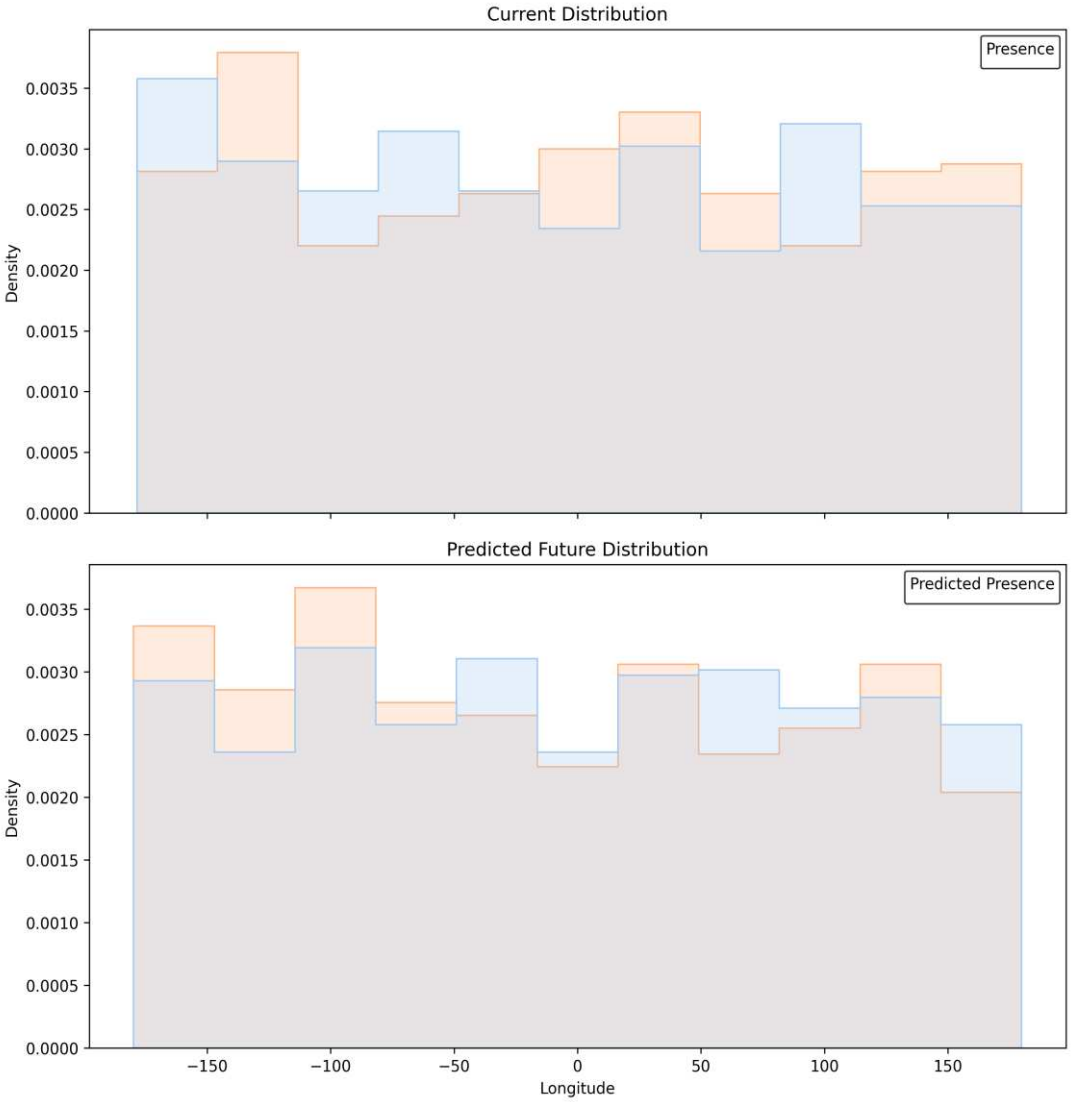

Fig. 3: Example Impact of Climate Change on Species Distribution

Forecasting the Spread of Invasive Species:

Invasive species can have detrimental eff cts on native ecosystems. XGBoost can help in predicting the potential spread of these species based on environmental conditions, aiding in early detection and management [16].

These applications showcase the potential of XGBoost in addressing some of the most pressing challenges in environmental ecology. The algorithm's ability to handle complex datasets and derive meaningful insights makes it a valuable asset for researchers and conservationists alike [17].

4. Benefits of Using XGBoost in Environmental Ecology

The application of XGBoost in environmental ecology is not just a matter of trend or convenience. The algorithm off rs several tangible benefits that make it particularly suited for ecological studies [6,18].

- High Accuracy and Predictive Power: One of the primary reasons for XGBoost's popularity is its exceptional accuracy. In many benchmark datasets and competitions, XGBoost has outperformed other machine learning algorithms, making it a reliable choice for ecological predictions [6,19].
- Ability to Handle Large Datasets: Environmental studies often involve vast amounts of data collected from various sources. XGBoost's scalability ensures that it can process these large datasets efficiently, providing timely and accurate insights [6,20].
- Robustness to Noise and Missing Data: Ecological data can be messy, with missing values or noise from various sources. XGBoost's built-in mechanisms for handling such inconsistencies make it a robust choice for real-world applications [21,22].
- Feature Importance: Understanding which variables signific ntly influence the model's predictions can be crucial in ecology. XGBoost provides a built-in feature importance metric, allowing researchers to identify key environmental factors driving ecological patterns [22,23].
- Flexibility: Whether it's predicting the spread of an invasive species, assessing habitat suitability, or modeling the impact of climate change, XGBoost's flexibility ensures it can be tailored to a wide range of ecological tasks [15,24].

The combination of accuracy, robustness, and flexibility makes XGBoost a powerful tool in the arsenal of environmental ecologists, enabling them to derive deeper insights and make more informed decisions.

5. Challenges and Limitations

While XGBoost off rs numerous advantages, it's essential to be aware of its limitations and challenges, especially when applied to environmental ecology [23]:

- Overfitt ng: Like many machine learning algorithms, XGBoost can sometimes fit the training data too closely, leading to poor generalization on new, unseen data. While its built-in regularization helps mitigate this, careful tuning and cross-validation are essential [25,26].
- Interpretability: Decision trees, in general, are considered interpretable models. However, when boosting hundreds or thousands of trees, as in XGBoost, the resulting model can become complex and harder to interpret. This can be a challenge when trying to understand ecological phenomena or when explaining results to stakeholders [26,27].

- Need for Domain Knowledge: While XGBoost can identify patterns in data, meaningful ecological interpretations require domain expertise. It's crucial to collaborate with ecologists to ensure that the model's predictions align with ecological theory and understanding [20,28].
- Computational Demands: Although XGBoost is optimized for performance, training on very large datasets or tuning hyperparameters can be computationally intensive, requiring adequate hardware resources [29,30].
- Data Quality: The adage "garbage in, garbage out" holds true for XGBoost. The quality of predictions is heavily dependent on the quality of the input data. Ensuring accurate and representative data collection is paramount [31,32].

Recognizing these challenges and limitations is crucial for the eff ctive application of XGBoost in environmental ecology. By being aware of potential pitfalls and working collaboratively with domain experts, researchers can harness the power of XGBoost while navigating its limitations.

6. Conclusion

XGBoost has undeniably carved a niche for itself in the realm of machine learning, and its applications in environmental ecology are a testament to its versatility and power. From predicting species distributions to modeling the impacts of climate change, XGBoost has provided researchers with a robust tool to tackle some of the most pressing ecological challenges of our time.

However, like any tool, its eff ctiveness is determined by how it's used. The challenges and limitations of XGBoost underscore the importance of a holistic approach, where domain knowledge, data quality, and algorithmic understanding come together. By leveraging the strengths of XGBoost and addressing its limitations, researchers and conservationists can push the boundaries of what's possible in environmental ecology.

In an era where data-driven decision-making is paramount, XGBoost stands as a beacon, illuminating the path forward for a more sustainable and harmonious coexistence with our environment.

References:

[1] Greener, J.G.; Kandathil, S.M.; Moffa , L.; Jones, D.T. A Guide to Machine Learning for Biologists. *Nat Rev Mol Cell Biol* 2022, *23*, 40–55, doi:10.1038/s41580-021-00407-0.

[2] Janiesch, C.; Zschech, P.; Heinrich, K. Machine Learning and Deep Learning. *Electronic Markets* 2021, *31*, 685–695, doi:10.1007/s12525-021-00475-2.

[3] Asselman, A.; Khaldi, M.; Aammou, S. Enhancing the Prediction of Student Performance Based on the Machine Learning XGBoost Algorithm. *Interactive Learning Environments* 2023, *31*, 3360–3379, doi:10.1080/10494820.2021.1928235.

[4] Li, Z. Extracting Spatial Effects from Machine Learning Model Using Local Interpretation Method: An Example of SHAP and XGBoost. *Comput Environ Urban Syst* 2022, *96*, 101845, doi:10.1016/j.compenvurbsys.2022.101845.

[5] Guan, G.; Liu, D.; Zhai, J. Factors Influencing Consumer Satisfaction of Fresh Produce E-Commerce in the Background of COVID-19—A Hybrid Approach Based on LDA-SEM-XGBoost. *Sustainability* 2022, *14*, 16392, doi:10.3390/su142416392.

[6] Kiangala, S.K.; Wang, Z. An Effective Adaptive Customization Framework for Small Manufacturing Plants Using Extreme Gradient Boosting-XGBoost and Random Forest Ensemble Learning Algorithms in an Industry 4.0 Environment. *Machine Learning with Applications* 2021, *4*, 100024, doi:10.1016/j.mlwa.2021.100024.

[7] Lartey, B.; Homaifar, A.; Girma, A.; Karimoddini, A.; Opoku, D. XGBoost: A Tree-Based Approach for Traffic Volume Prediction. In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC); IEEE, October 17 2021; pp. 1280–1286.

[8] Prakash, A.; Thangaraj, J.; Roy, S.; Srivastav, S.; Mishra, J.K. Model-Aware XGBoost Method Towards Optimum Performance of Flexible Distributed Raman Amplifier. *IEEE Photonics J* 2023, *15*, 1–10, doi:10.1109/JPHOT.2023.3286272.

[9] Lei, Y.; Jiang, W.; Jiang, A.; Zhu, Y.; Niu, H.; Zhang, S. Fault Diagnosis Method for Hydraulic Directional Valves Integrating PCA and XGBoost. *Processes* 2019, *7*, 589, doi:10.3390/pr7090589.

[10] Bhati, B.S.; Chugh, G.; Al-Turjman, F.; Bhati, N.S. An Improved Ensemble Based Intrusion Detection Technique Using <scp>XGBoost</Scp>. *Transactions on Emerging Telecommunications Technologies* 2021, *32*, doi:10.1002/ett.4076.

[11] Valavi, R.; Guillera-Arroita, G.; Lahoz-Monfort, J.J.; Elith, J. Predictive Performance of Presence-only Species Distribution Models: A Benchmark Study with Reproducible Code. *Ecol Monogr* 2022, *92*, doi:10.1002/ecm.1486.

[12] Cha, Y.; Shin, J.; Go, B.; Lee, D.-S.; Kim, Y.; Kim, T.; Park, Y.-S. An Interpretable Machine Learning Method for Supporting Ecosystem Management: Application to Species Distribution Models of Freshwater Macroinvertebrates. *J Environ Manage* 2021, *291*, 112719, doi:10.1016/j.jenvman.2021.112719.

[13] Wieland, R.; Kuhls, K.; Lentz, H.H.K.; Conraths, F.; Kampen, H.; Werner, D. Combined Climate and Regional Mosquito Habitat Model Based on Machine Learning. *Ecol Modell* 2021, *452*, 109594, doi:10.1016/j.ecolmodel.2021.109594.

[14] Ghafarian, F.; Wieland, R.; Lüttschwager, D.; Nendel, C. Application of Extreme Gradient Boosting and Shapley Additive Explanations to Predict Temperature Regimes inside Forests from Standard Open-Field Meteorological Data. *Environmental Modelling & Software* 2022, *156*, 105466, doi:10.1016/j.envsoft.2022.105466.

[15] Liu, X.; Chen, X.; Potoglou, D.; Tian, M.; Fu, Y. Travel Impedance, the Built Environment, and Customized-Bus Ridership: A Stop-to-Stop Level Analysis. *Transp Res D Transp Environ* 2023, *122*, 103889, doi:10.1016/j.trd.2023.103889.

[16] Farooq, Z.; Rocklöv, J.; Wallin, J.; Abiri, N.; Sewe, M.O.; Sjödin, H.; Semenza, J.C. Artificial Intelligence to Predict West Nile Virus Outbreaks with Eco-Climatic Drivers. *The Lancet Regional Health - Europe* 2022, *17*, 100370, doi:10.1016/j.lanepe.2022.100370.

[17] Bergamo, T.F.; de Lima, R.S.; Kull, T.; Ward, R.D.; Sepp, K.; Villoslada, M. From UAV to PlanetScope: Upscaling Fractional Cover of an Invasive Species Rosa Rugosa. *J Environ Manage* 2023, *336*, 117693, doi:10.1016/j.jenvman.2023.117693.

[18] Wang, L.; Zhao, C.; Liu, X.; Chen, X.; Li, C.; Wang, T.; Wu, J.; Zhang, Y. Non-Linear Effects of the Built Environment and Social Environment on Bus Use among Older Adults in China: An Application of the XGBoost Model. *Int J Environ Res Public Health* 2021, *18*, 9592, doi:10.3390/ijerph18189592.

[19] Yang, Y.; Wang, K.; Yuan, Z.; Liu, D. Predicting Freeway Traffic Crash Severity Using XGBoost-Bayesian Network Model with Consideration of Features Interaction. *J Adv Transp* 2022, *2022*, 1–16, doi:10.1155/2022/4257865.

[20] Henriques, J.; Caldeira, F.; Cruz, T.; Simões, P. Combining K-Means and XGBoost Models for Anomaly Detection Using Log Datasets. *Electronics (Basel)* 2020, *9*, 1164, doi:10.3390/electronics9071164.

[21] Hu, X.; Jia, H.; Zhang, Y.; Deng, Y. An Open-Circuit Faults Diagnosis Method for MMC Based on Extreme Gradient Boosting. *IEEE Transactions on Industrial Electronics* 2023, *70*, 6239–6249, doi:10.1109/TIE.2022.3194584.

[22] Muyama, L.; Neuraz, A.; Coulet, A. Extracting Diagnosis Pathways from Electronic Health Records Using Deep Reinforcement Learning. *arXiv preprint arXiv:2305.06295* 2023.

[23] Shi, C.; Wang, Y. Development of Subsurface Geological Cross-Section from Limited Site-Specific Boreholes and Prior Geological Knowledge Using Iterative Convolution XGBoost. *Journal of Geotechnical and Geoenvironmental Engineering* 2021, *147*, 04021082.

[24] Ren, X.; Mi, Z.; Georgopoulos, P.G. Comparison of Machine Learning and Land Use Regression for Fine Scale Spatiotemporal Estimation of Ambient Air Pollution: Modeling Ozone Concentrations across the Contiguous United States. *Environ Int* 2020, *142*, 105827, doi:10.1016/j.envint.2020.105827.

[25] Thongsuwan, S.; Jaiyen, S.; Padcharoen, A.; Agarwal, P. ConvXGB: A New Deep Learning Model for Classification Problems Based on CNN and XGBoost. *Nuclear Engineering and Technology* 2021, *53*, 522–531, doi:10.1016/j.net.2020.04.008.

[26] Abdullah, T.A.A.; Zahid, M.S.M.; Ali, W. A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. *Symmetry (Basel)* 2021, *13*, 2439, doi:10.3390/sym13122439.

[27] Uddin, M.N.; Li, L.-Z.; Deng, B.-Y.; Ye, J. Interpretable XGBoost–SHAP Machine Learning Technique to Predict the Compressive Strength of Environment-Friendly Rice Husk Ash Concrete. *Innovative Infrastructure Solutions* 2023, *8*, 147, doi:10.1007/s41062-023-01122-9.

[28] Guo, Z.; Ding, N.; Zhai, M.; Zhang, Z.; Li, Z. Leveraging Domain Knowledge to Improve Depression Detection on Chinese Social Media. *IEEE Trans Comput Soc Syst* 2023, *10*, 1528–1536, doi:10.1109/TCSS.2023.3267183.

[29] Zhao, X.; Li, Q.; Xue, W.; Zhao, Y.; Zhao, H.; Guo, S. Research on Ultra-Short-Term Load Forecasting Based on Real-Time Electricity Price and Window-Based XGBoost Model. *Energies (Basel)* 2022, *15*, 7367, doi:10.3390/en15197367.

[30] Gajjar, A.; Kashyap, P.; Aysu, A.; Franzon, P.; Dey, S.; Cheng, C. FAXID: FPGA-Accelerated XGBoost Inference for Data Centers Using HLS. In Proceedings of the 2022 IEEE 30th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM); IEEE, May 15 2022; pp. 1–9.

[31] Jing, R.; Tian, H.; Li, Y.; Zhang, X.; Zheng, X.; Zhang, Z.; Zeng, D. Improving the Data Quality for Credit Card Fraud Detection. In Proceedings of the 2020 IEEE International Conference on Intelligence and Security Informatics (ISI); IEEE, November 9 2020; pp. 1–6.

[32] Cao, D.; Ma, Y.; Sun, L.; Gao, L. Fast Observation Simulation Method Based on XGBoost for Visible Bands over the Ocean Surface under Clear-Sky Conditions. *Remote Sensing Letters* 2021, *12*, 674–683, doi:10.1080/2150704X.2021.1925371.

