KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)
Terakreditasi Nomor 204/E/KPT/2022 | Vol. 5, No. 1, Januari (2024), pp. 189-194

Enhancing Riverine Water Quality Prediction: The Application of Variational Autoencoders for Robust Data Augmentation in Environmental Science

Gregorius Airlangga
Information System Department, Universitas Katolik Indonesia Atma Jaya, Indonesia
E-mail: gregorius.airlangga@atmajaya.ac.id

Abstract

In this study, we present a comprehensive approach to address a critical challenge in environmental science: the accurate prediction of dissolved oxygen (DO) levels in river ecosystems. Leveraging advanced machine learning techniques, particularly Variational Autoencoders (VAEs), our research aims to overcome the limitations posed by sparse and incomplete environmental datasets. We meticulously curated a dataset from multiple water monitoring stations, capturing key indicators such as DO, ammonium ions, nitrites, nitrates, and biochemical oxygen demand. Following data standardization and quality assessment, we implemented a RandomForestRegressor to ascertain feature importance, utilizing GridSearchCV and RandomizedSearchCV for model optimization. This allowed for precise feature selection to inform the predictive model. Anomaly detection was performed using One-Class SVM and Isolation Forest methodologies, essential for purifying the dataset by removing outliers. Subsequently, VAEs were applied to augment the data, synthesizing new data points that were statistically coherent with the original set, thus enriching the dataset and potentially unveiling concealed patterns. The augmented data's impact was evaluated through a RandomForestRegressor model, comparing RMSE scores before and after data augmentation, revealing a notable improvement in predictive accuracy with the lowest RMSE observed for the model utilizing VAE-generated data. This underscores the VAE's value in enhancing the model's performance, indicating that the synthetic data provided additional variability and complexity that aided the model's learning process. Our findings indicate that integrating sophisticated data augmentation techniques like VAEs can significantly enhance the quality of environmental datasets and the accuracy of predictive models.

Keywords: *Environmental, Data Science, Data Augmentation, Anomaly Detection*

1. Introduction

The pursuit of precise water quality assessment in river ecosystems represents a longstanding challenge in environmental science [1]–[3]. Traditional methods have primarily focused on direct measurements and statistical modeling to predict key indicators, notably dissolved oxygen (DO) levels [4]–[6]. While these methods are foundational, they face challenges in addressing the sparse and often incomplete nature of environmental datasets, which are characterized by complex, non-linear ecological interactions [7]–[9]. The evolution of environmental monitoring has seen the integration of advanced statistical models and machine learning techniques, aiming to enhance the accuracy of predictions [10]–[12]. However, a significant challenge remains: the development of robust predictive models is often hindered by the intrinsic limitations of available data [13]–[15]. Addressing this pivotal gap, our study introduces an innovative application of Variational Autoencoders (VAEs) for data generation, representing a paradigm shift in the approach to environmental data analysis. VAEs uniquely combine deep learning with Bayesian inference to synthesize new data points that maintain statistical consistency with the original dataset [16]–[18]. This is particularly

ISSN: 2720-992X

advantageous in environmental science, where the complexity of ecosystems often surpasses the scope of conventional data collection methods [19]. By employing VAEs, our study aims to overcome the constraints of data scarcity and quality, enriching datasets and revealing latent patterns that might be obscured in smaller data samples.

Our research meticulously develops a VAE architecture specifically designed for environmental data, optimizing the model through advanced hyperparameter tuning and validation methods like GridSearchCV and RandomizedSearchCV. The VAE's encoder component efficiently compresses high-dimensional input data into a compact latent space, capturing key features essential for predicting DO levels. Subsequently, the decoder component reconstructs the input data from this latent representation, generating new, synthetic data points that reflect the intricate relationships found in the original dataset. A cornerstone of our study is the thorough evaluation of the synthetic data produced by the VAE. Using a RandomForestRegressor model, we conduct a detailed comparison of the predictive performance before and after data augmentation, with Root Mean Squared Error (RMSE) serving as the primary metric for accuracy assessment. Additionally, we implement cosine similarity measures to evaluate the alignment between the original and generated datasets, ensuring that the synthetic data accurately reflects the environmental variables' true characteristics.

This research marks a contribution to environmental data science, underscoring the potential of VAEs to transform the field of ecological modeling and prediction. By enhancing the depth and reliability of datasets, our approach holds promise for improving river ecosystem management, offering a pathway for more informed and effective environmental stewardship. Structured to offer a thorough overview of the methodology, the paper progresses to a detailed presentation of the results derived from the applied VAE techniques for data generation. The subsequent analysis delves into the implications of these findings for environmental monitoring and predictive modeling. The conclusion synthesizes the key discoveries and proposes future research directions. This systematic approach ensures that the paper comprehensively addresses the study's aims while contributing valuable insights to the field of environmental monitoring.

2. Reseach Methodology

Our study adopts a multifaceted methodology to enhance the predictive accuracy of dissolved oxygen (DO) levels in river ecosystems, a crucial parameter in assessing water quality. This approach combines advanced data preprocessing, feature analysis, anomaly detection, and state-of-the-art machine learning techniques, including the novel application of Variational Autoencoders (VAEs) for data augmentation. The dataset has five indicators that are measured at 8 stations of the state water monitoring system. Indicators of river water quality in this dataset are:

a) O2(i): Dissolved oxygen (O2) is measured in mgO2/cub. dm (i.e. milligrams of oxygen (O2) in the cubic decimeter).
b) NH4(i): Ammonium ions (NH4) concentration is measured in mg/cub. dm (i.e. milligrams in the cubic decimeter).
c) NO2(i): Nitrite ions (NO2) concentration is measured in mg/cub. dm (i.e. milligrams in the cubic decimeter).
d) NO3(i): Nitrate ions (NO3) concentration is measured in mg/cub. dm (i.e. milligrams in the cubic decimeter).
e) BOD5(i): Biochemical oxygen demand, which is determined in 5 days ("BOD5" or "BOD"). BOD5 is measured in MgO/cub. dm (i.e. milligrams of oxygen in cubic decimeters).

2.1. Data Collection and Preparation

We utilize three primary datasets: 'train.csv' for training, 'test.csv' for validation, and 'sample_submission.csv' as a template for outputs. Each dataset undergoes a standardization process where columns are renamed for consistency, thereby facilitating easier data manipulation and analysis. Furthermore, we utilize tools such as PrettyTable and missingno (msno), we conduct an exhaustive assessment of data quality. This involves analyzing data types, identifying missing values, and determining the percentage of data missing in each column. This rigorous analysis allows us to understand the extent of data completeness and plan appropriate preprocessing steps.

2.2. Feature Importance and Selection

We employ a RandomForestRegressor to evaluate the importance of various features in predicting DO levels. This model is optimized using hyperparameter tuning methods such as GridSearchCV and RandomizedSearchCV. The performance is evaluated through KFold cross-validation, focusing on metrics like Mean Absolute Error (MAE) to assess the model's predictive power. In addition, The most influential features are identified and iteratively added to the model. At each step, the model's performance is assessed using the Root Mean Squared Error (RMSE). This process helps in pinpointing the optimal set of features that contribute significantly to prediction accuracy.

2.3. Anomaly Detection and Data Augmentation

Methods like One-Class SVM and Isolation Forest are deployed to detect and remove outliers from the dataset. This step is crucial as outliers can significantly skew model predictions and affect the overall data quality. We design a VAE architecture specifically tailored for environmental data. The VAE consists of an encoder network, which compresses high-dimensional data into a latent space, and a decoder network, which reconstructs data from this latent representation. The model's loss function includes both the reconstruction loss and KL divergence, ensuring that the synthetic data generated is realistic and representative of the original dataset. Furthermore, the synthetic data produced by the VAE is carefully integrated with the original dataset. We then employ a RandomForestRegressor to evaluate the augmented dataset's quality. The RMSE metric is used to quantify the improvement in predictive accuracy. Additionally, cosine similarity analysis ensures that the synthetic data closely mirrors the characteristics of the original dataset.

2.4. Model Training and Evaluation

The effectiveness of data augmentation is assessed by comparing the models' performance, pre- and post-augmentation. This comparison, focusing on RMSE, helps determine the impact of VAE-generated data on improving model accuracy. Furthermore, the final model selection is based on a combination of accuracy and generalization. The model demonstrating the highest predictive accuracy with the augmented dataset is chosen as the primary tool for predicting DO levels in river ecosystems.

3. Results and Discussion

As presented on table 1, the results indicate that data augmentation using VAEs has a positive impact on the predictive accuracy of the Random Forest model, as evidenced by the lowest RMSE score. This suggests that the additional synthetic data provided by the VAE can enhance the model's training process, potentially by introducing a broader variety of examples that help the model generalize better to unseen data. The RMSE values across different methods highlight the importance of a comprehensive approach that includes both anomaly detection and data augmentation in environmental data analysis. The use of sophisticated algorithms like VAEs in conjunction with robust

predictive models like Random Forest can lead to more accurate environmental monitoring tools, which are crucial for effective management and conservation efforts.

The VAE has an RMSE of 1.141, which indicates a moderate level of prediction error. As a generative model, the VAE's primary function in this context is likely data augmentation. Its performance suggests that the synthetic data it generated was reasonably effective in improving the model's accuracy. LOF, an algorithm used for identifying outliers in data, shows a slightly higher RMSE of 1.152 compared to the VAE. This could imply that while LOF is effective in detecting anomalies, the method might either remove some useful information along with the outliers or not detect all the anomalies, which slightly decreases the predictive accuracy.

With an RMSE of 1.284, PCA, a dimensionality reduction technique, appears to be the least effective among the methods listed. This could be due to PCA's tendency to preserve variance as opposed to directly improving predictive accuracy. It suggests that the principal components retained may not capture all the variables critical for predicting DO levels. The standalone Random Forest model yields an RMSE of 1.142, which is slightly better than the VAE and LOF but not by a large margin. This indicates that the Random Forest model on its own is relatively effective in capturing the patterns in the data necessary for making accurate predictions. The Random Forest model using data augmented by the VAE shows the lowest RMSE of 1.068. This is a significant improvement over the standalone Random Forest model and suggests that the synthetic data added value to the model, helping to capture more of the underlying data distribution and improve the accuracy of DO level predictions.

Table 1. Comparison Result

Metric/Method	VAE	Local Outlier Factor	PCA	Random Forest Regressor	Random Forest Regressor with generated data from VAE
RMSE	1.141	1.152	1.284	1.142	1.068

4. Conclusion

In conclusion, our investigation into the enhancement of predictive models for river water quality, specifically dissolved oxygen (DO) levels, has yielded insightful findings. We embarked on this research endeavor to address the critical gap in environmental data science—particularly the limitations posed by sparse and incomplete datasets—and our results underscore the efficacy of employing advanced machine learning techniques for this purpose. Throughout the study, we adhered to a structured methodology that encompassed data preprocessing, feature importance analysis, anomaly detection, and the application of machine learning models. Notably, the integration of Variational Autoencoders (VAEs) for data augmentation stood out as a pivotal innovation, enhancing our dataset and revealing latent patterns that traditional methods might overlook.

The results from our methodical approach were quantified using the Root Mean Squared Error (RMSE) as the primary metric for accuracy. The application of VAEs for generating synthetic data proved to be particularly effective, as evidenced by the improved RMSE scores when this data was used in conjunction with a RandomForest Regressor model. The VAE-augmented model outperformed other methods, including PCA and Local Outlier Factor, and even the standalone RandomForest model, emphasizing the value of VAE-generated data in enhancing model accuracy. Our research has not only demonstrated the potential of VAEs in creating robust predictive models for environmental monitoring but also set a benchmark for future studies in this field.

By improving the depth and quality of environmental datasets, we contribute to the advancement of ecological conservation efforts and the development of effective river ecosystem management strategies. The implications of our findings are significant, paving

the way for the adoption of sophisticated data augmentation techniques in environmental data analysis and beyond. Looking ahead, we advocate for continued exploration of VAEs and other generative models, foreseeing their broader application across various domains of environmental science. The integration of such models with an array of machine learning algorithms can potentially lead to groundbreaking advancements in predictive accuracy, model robustness, and ultimately, the sustainability of natural ecosystems.

References

[1] Y. Huang *et al.*, "Forward-looking roadmaps for long-term continuous water quality monitoring: bottlenecks, innovations, and prospects in a critical review," *Environ. Sci. \& Technol.*, vol. 56, no. 9, pp. 5334–5354, 2022.

[2] S. Khullar and N. Singh, "Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation," *Environ. Sci. Pollut. Res.*, vol. 29, no. 9, pp. 12875–12889, 2022.

[3] L. M. Kuehne *et al.*, "The future of global river health monitoring," *PLOS Water*, vol. 2, no. 9, p. e0000101, 2023.

[4] C. Xu, X. Chen, and L. Zhang, "Predicting river dissolved oxygen time series based on stand-alone models and hybrid wavelet-based models," *J. Environ. Manage.*, vol. 295, p. 113085, 2021.

[5] Y. Liu, Q. Zhang, L. Song, and Y. Chen, "Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction," *Comput. Electron. Agric.*, vol. 165, p. 104964, 2019.

[6] H. Guo *et al.*, "A generalized machine learning approach for dissolved oxygen estimation at multiple spatiotemporal scales using remote sensing," *Environ. Pollut.*, vol. 288, p. 117734, 2021.

[7] S. Luo *et al.*, "FREE: The Foundational Semantic Recognition for Modeling Environmental Ecosystems," *arXiv Prepr. arXiv2311.10255*, 2023.

[8] H. Li, C. Qin, W. He, F. Sun, and P. Du, "Improved predictive performance of cyanobacterial blooms using a hybrid statistical and deep-learning method," *Environ. Res. Lett.*, vol. 16, no. 12, p. 124045, 2021.

[9] M. Sonnewald, R. Lguensat, D. C. Jones, P. D. Dueben, J. Brajard, and V. Balaji, "Bridging observations, theory and numerical simulation of the ocean using machine learning," *Environ. Res. Lett.*, vol. 16, no. 7, p. 73008, 2021.

[10] S. Zhong *et al.*, "Machine learning: new ideas and tools in environmental science and engineering," *Environ. Sci. \& Technol.*, vol. 55, no. 19, pp. 12741–12754, 2021.

[11] D. Iskandaryan, F. Ramos, and S. Trilles, "Air quality prediction in smart cities using machine learning technologies based on sensor data: a review," *Appl. Sci.*, vol. 10, no. 7, p. 2401, 2020.

[12] W. Wei, O. Ramalho, L. Malingre, S. Sivanantham, J. C. Little, and C. Mandin, "Machine learning and statistical models for predicting indoor air quality," *Indoor Air*, vol. 29, no. 5, pp. 704–726, 2019.

[13] W. Wu, R. Emerton, Q. Duan, A. W. Wood, F. Wetterhall, and D. E. Robertson, "Ensemble flood forecasting: Current status and future opportunities," *Wiley Interdiscip. Rev. Water*, vol. 7, no. 3, p. e1432, 2020.

[14] A. J. Lopatkin and J. J. Collins, "Predictive biology: modelling, understanding and harnessing microbial complexity," *Nat. Rev. Microbiol.*, vol. 18, no. 9, pp. 507–520, 2020.

[15] F. Couvreux *et al.*, "Process-based climate model development harnessing machine learning: I. A calibration tool for parameterization improvement," *J. Adv. Model. Earth Syst.*, vol. 13, no. 3, p. e2020MS002217, 2021.

[16] J. Houston, F. G. Glavin, and M. G. Madden, "Robust classification of high-

dimensional spectroscopy data using deep learning and data synthesis," *J. Chem. Inf. Model.*, vol. 60, no. 4, pp. 1936–1954, 2020.

[17] R. Tang, "Some advances in Bayesian inference and generative modeling," University of Illinois at Urbana-Champaign, 2023.

[18] T. Glüsenkamp, "Unifying supervised learning and VAEs--automating statistical inference in (astro-) particle physics with amortized conditional normalizing flows," *arXiv Prepr. arXiv2008.05825*, 2020.

[19] A. De Vos, R. Biggs, and R. Preiser, "Methods for understanding social-ecological systems: a review of place-based studies," *Ecol. Soc.*, vol. 24, no. 4, 2019.