Lecture Notes in Networks and Systems 860

Miroslav Trajanović
Nenad Filipović
Milan Zdravković *Editors*

Disruptive Information Technologies for a Smart Society

Proceedings of the 14th International Conference on Information Society and Technology (ICIST)

Springer

Lecture Notes in Networks and Systems 860

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Miroslav Trajanović · Nenad Filipović ·
Milan Zdravković
Editors

Disruptive Information Technologies for a Smart Society

Proceedings of the 14th International Conference on Information Society and Technology (ICIST)

Springer

Editors
Miroslav Trajanović
Faculty of Mechanical Engineering
University of Niš
Niš, Serbia

Nenad Filipović
Faculty of Engineering
University of Kragujevac
Kragujevac, Serbia

Milan Zdravković
Faculty of Mechanical Engineering
University of Niš
Niš, Serbia

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-71418-4 ISBN 978-3-031-71419-1 (eBook)
https://doi.org/10.1007/978-3-031-71419-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

In the era of unprecedently rapid technological advancement, the integration of innovative information technologies into various sectors has become pivotal in shaping a smarter and more interconnected society. The 14th International Conference on Information Society and Technology has once again brought an insight into these technologies through a collection of groundbreaking research, compiled in this volume. This book presents a comprehensive overview of the latest advancements and practical applications across multiple domains.

The volume is organized into five distinct sections, each highlighting a critical area where information technologies are driving significant transformations: AI-enhanced Industry, Digitalization in Health, Well-being, and Sport, Enterprise Information Systems of the Future, Large Language Models, and Security and Safety.

The first section explores the transformative potential of artificial intelligence in industrial contexts. Research in this section focuses on the application of advanced AI techniques to optimize processes, enhance efficiency, and predict outcomes in various industrial systems. The integration of machine learning models and decision support systems exemplifies the innovative approaches being developed to address complex industrial challenges. These advancements underscore the crucial role of AI in modernizing industries, making them more responsive and adaptive to changing conditions. By leveraging AI, industries can achieve significant improvements in productivity, sustainability, and operational efficiency, paving the way for a new era of industrial innovation.

Digitalization is revolutionizing the health and well-being sectors, offering novel opportunities for improving patient care, diagnostics, and sports performance analysis. This section presents research that leverages digital technologies to enhance health outcomes and well-being. Studies such as those on marker-less motion analysis, algorithm-based impact assessments, and the application of computer vision models in sport highlight the potential of digital solutions to transform healthcare delivery and sports analytics, ultimately improving quality of life.

The evolution of information systems is a key focus in the third section, where innovative approaches are being developed to enhance business processes and improve efficiency. Research in this section addresses the integration of AI into enterprise systems, automation of testing processes, and the development of advanced educational tools. These contributions signify a leap toward the next generation of enterprise solutions, characterized by increased automation and enhanced process awareness, reflecting a forward-looking vision for enterprise systems that are more intelligent and adaptive. By harnessing the power of AI and automation, businesses can streamline operations, reduce costs, and improve decision-making processes, ultimately achieving greater competitiveness in the global market.

Large language models (LLMs) have emerged as a significant breakthrough in natural language processing, offering new possibilities for understanding and generating human language. This section discusses the capabilities and applications of LLMs,

including comparative analyses, semantic exploration, and emotion detection. These studies demonstrate the broad applicability and transformative power of LLMs. They push the boundaries of language technologies, providing insights into how LLMs can be leveraged to improve communication, education, and information retrieval.

The final section addresses the critical issues of security and safety in the digital age. As technology advances, ensuring the reliability and security of systems and data becomes increasingly important. This section presents research focused on developing robust security measures, risk assessment platforms, and methodologies for safeguarding digital assets. The innovative approaches discussed here also aim to mitigate risks, enhance structural safety, and ensure the secure configuration of web applications. These contributions highlight the ongoing efforts to build a safer and more secure digital environment.

In conclusion, "Disruptive Information Technologies for a Smart Society" provides a comprehensive overview of the cutting-edge research presented at the 14th International Conference on Information Society and Technology. The breadth and depth of the research compiled in this volume underscores the pivotal role of information technologies in driving societal progress. This book not only serves as a repository of advanced knowledge but also as an inspiration for future research and development in the field of information technology. As editors, we extend our gratitude to the authors and reviewers whose contributions have made this publication possible.

Miroslav Trajanović
Nenad Filipović
Milan Zdravković

Contents

Digitalisation in Health, Well-being and Sport

Enterprise Information Systems of the future

Large Language Models

AI-enhanced Industry

Machine Learning Model for Prediction of Indicative Water Parameters on the Danube River Based on Satellite Data

Velibor Ilić(✉), Milan Stojković, Zorica Dodevska, and Slobodan Ilić

The Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, Serbia
velibor.ilic@ivi.ac.rs

Abstract. This study introduces a novel machine-learning approach using Sentinel-2 satellite images for water quality assessment in the Danube River. Utilizing a deep neural network to build a model that integrates multispectral satellite data with in-situ measurements, the research provides a comprehensive analysis with augmented data. It demonstrates high predictive accuracy for significant water quality indicators for the Danube River. The R2-result exceeds 0.98 for water temperature, electrical conductivity, and dissolved oxygen, while slightly less precision is achieved for chemical oxygen demand. Our method represents a scalable, efficient improvement of traditional assessment techniques, emphasizing the synergy of remote sensing and machine learning. It significantly advances monitoring water quality parameters in hydrology stations on river flows near urban environments with the possibility of implementing it in other locations of interest. In addition, our approach that uses filtering masks is adaptable to different in-land water surface satellite-based precise detections.

Keywords: machine learning · deep neural network · satellite data · Sentinel-2 · water quality parameters · Danube River

1 Introduction

In recent years, advances in satellite technology have enabled new ways to monitor the environment, especially in the assessment of water quality parameters on a global scale [1–7]. Integrating machine learning (ML) with satellite data [8–11] represents a state-of-the-art approach to predicting significant water quality indicators, offering a comprehensive and effective method for environmental management and policy-making. Moreover, ML techniques are increasingly employed for spatial interpolation, enhancing the prediction of environmental variables across large, unsampled areas by leveraging patterns detected in satellite imagery [12–14]. This paper introduces a new ML model, based on filtering masks for more precise detection of in-land water surfaces, designed to use satellite data to accurately predict critical water parameters such as water temperature (T), electrical conductivity (EC), dissolved oxygen (DO), and chemical oxygen demand (COD).

M. Trajanović et al. (Eds.): ICIST 2024, LNNS 860, pp. 3–11, 2024.
https://doi.org/10.1007/978-3-031-71419-1_1

Using large amounts of data available from Earth observation satellites, the proposed model aims to provide insight into water quality trends, enabling timely and effective responses to environmental challenges. Our research bridges the gap of predicting river parameters measured at one point (i.e., at the hydrology station) to another where measurements are not available in appropriate frequencies, offering a novel water resource management and conservation framework that efficiently combines remote sensing technology and machine learning techniques. We validated the model on hydrology stations (HS) located on the Dunav River at the two biggest cities in Serbia, Belgrade (HS Zemun) and Novi Sad (HS Varadin Bridge).

2 Research Questions

Table 1 represents an overview of the research questions essential for this study.

Table 1. An overview of research questions.

No	Research question	The purpose of the research question
1	What ML model can effectively predict indicative water parameters, such as T, EC, DO, and COD, from satellite data?	This question addresses the core objective of the research: developing an ML model tailored to interpret satellite imagery for water quality assessment
2	How does the performance of the developed model compare with traditional water quality monitoring methods?	This question explores the preprocessing steps required to transform satellite data into a format suitable for ML analysis
3	How does the performance of the developed model compare with traditional water quality monitoring methods?	This question aims to evaluate the effectiveness, efficiency, and scalability of the ML model against conventional methods
4	What are the limitations of using satellite data and ML in water quality prediction, and how can these be mitigated?	This question helps in understanding the constraints of this approach and proposing solutions or improvements
5	What are the broader implications of this research for environmental monitoring?	This question brings the opportunity to augment the available data of environmental critical points via satellite imagery and improves the quality of environmental decision-making
6	How can the model be adapted or expanded to other environmental monitoring applications?	This question explores the wider impact of the research and its potential applications beyond the scope of water quality monitoring

3 Methodology

Sentinel-2 satellite is equipped with a Multispectral Imager (MSI). This sensor has 13 spectral bands with pixel sizes ranging from 10 to 60 m. It has a 10-m resolution in its blue (B2), green (B3), red (B4), and near-infrared (B8) channels. The ground sampling distance for its red edge (B5), near-infrared NIR (B6, B7, and B8A), and short-wave infrared SWIR (B11 and B12) is 20 m. Finally, its coastal aerosol (B1) and cirrus band (B10) have pixel sizes of 60 m.

To prepare the data for prediction of water quality indicators at specific coordinates (longitude, latitude), we first download satellite data covering all 13 bands for a 3 × 3 km area, ensuring that the specified coordinates are at the center. After this step, a filtering mask is applied to the data to retain only values corresponding to measurements taken on the water surface. We then calculate the values for each of the 13 bands for these filtered points (water surface). As a result, we obtain data for all 13 bands for each point. Additionally, we assign a 'day_of_year' variable to each point, which is determined by the date of the data. These 14 values (13 average values of the bands and 'day_of_year') represent the input for the ML model. Satellite data were collected at locations where there are two measuring stations, HS Varadin Bridge (Novi Sad) and HS Zemun (Belgrade), as shown in Fig. 1.

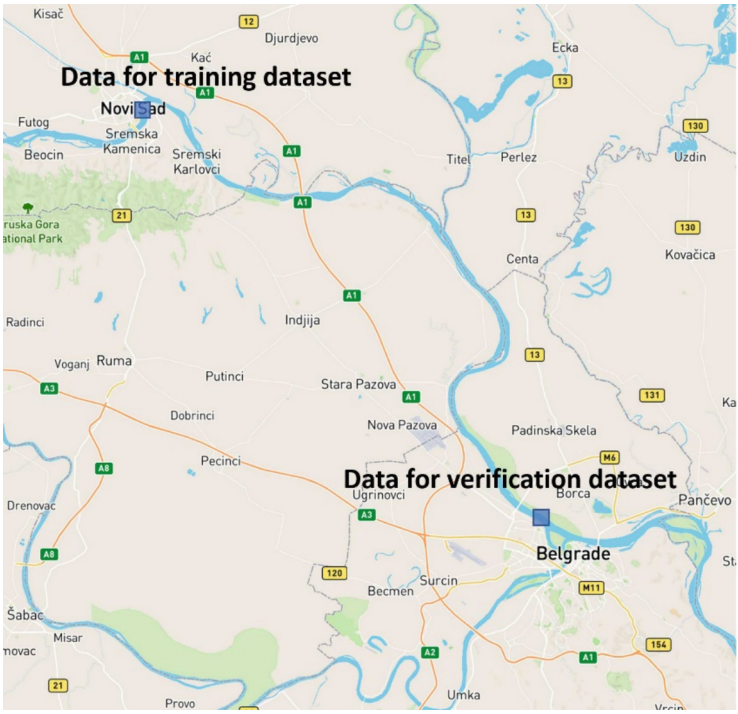

Fig. 1. Map with two hydrology stations at Belgrade (HS Zemun) and Novi Sad (HS Varadin Bridge).

During the summer, water surfaces are most distinctly identifiable at Band7 values. Figure 2 illustrates these values within a 3 × 3 km area at the coordinate (44.85 Latitude, 20.41 Longitude) for January, and Fig. 3 July. The contrast between land and water values is particularly evident in the July readings. Masks for differentiating measurements over water and land can be developed using Band7 data from Sentinel, especially in July, where water surfaces exhibit lower values compared to higher land values. Figure 4 shows only filtered measured values over water surfaces. These data filtering masks, designed in this manner, are applied across all other bands and throughout the entire year.

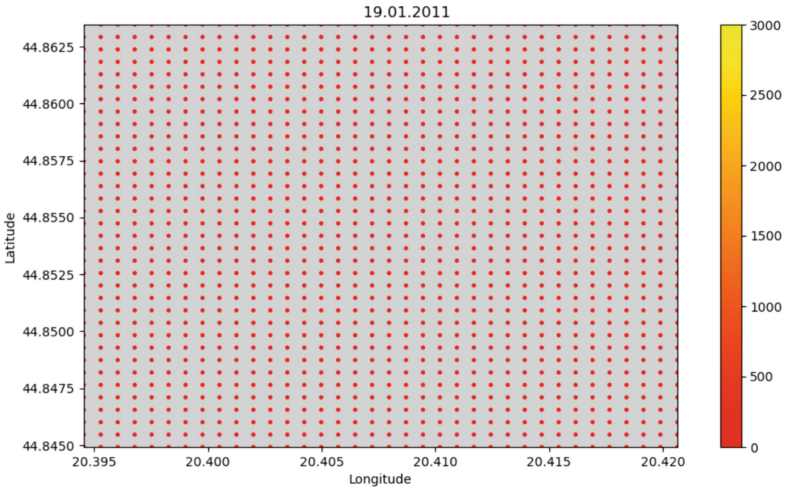

Fig. 2. Values of Band7 at area 3 × 3 km at the coordinate 44.85 Lat, 20.41 Long for January.

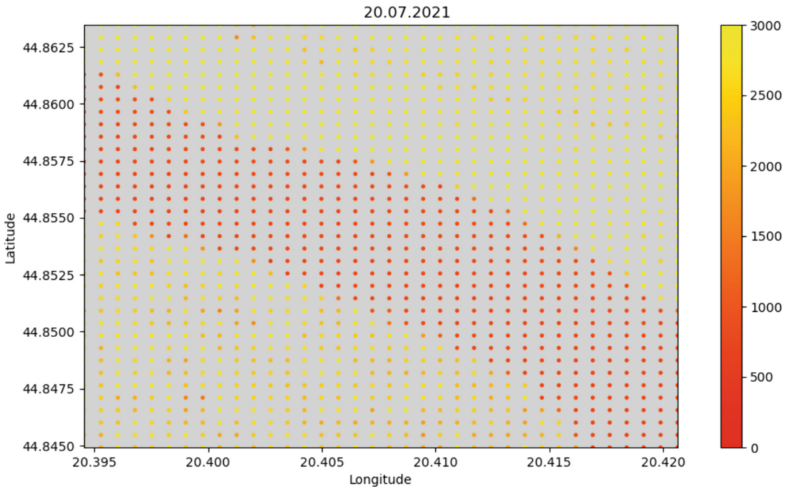

Fig. 3. Values of Band7 at area 3 × 3 km at the coordinate 44.85 Lat, 20.41 Long for July.

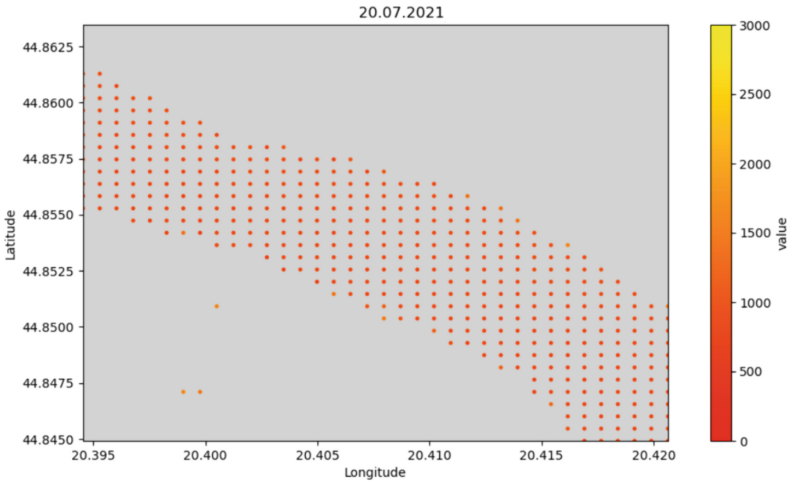

Fig. 4. Filtering masks for water surfaces.

Historical data about the Danube River's parameters, including T, EC, DO, and COD, have been available at the HS Varadin Bridge with daily frequency since October 2012. The data are collected from the Republic Hydrometeorological Service of Serbia (RHMS) and the Serbian Environmental Agency (SEPA) annual reports available on websites [15–18].

Additionally, data collection from the Sentinel-2 satellite started in July 2015. To train our ML model, we use both the satellite data and the measured historical data. The dataset contains information from July 2015 to December 2022. However, since the Sentinel-2 does not provide data each day, any measurements corresponding to dates without satellite data have been excluded from the dataset.

4 Solution/Discussion

The dataset is structured to provide input and output data for a machine learning model. Each row in the data set contains 18 values. The first 14 values are input values used as the input of the ML model: 13 average values of Sentinel Bands, and the 14th value is 'day_of_year'. The remaining four values represent water quality indicators: T (°C), EC (μS/cm), DO (mg/l), and COD (mg/l), values that the model should learn to predict at the output of the network. The dataset created in this way contains 2065 rows of data.

A deep neural network (DNN) model was used in this project, (Fig. 5). It contains 14 inputs (13 range values from the Sentinel-2 satellite and the 'day_of_year' value), three hidden layers, and four outputs (T, EC, DO, and COD). The first, second, and third hidden layers contain 256, 128, and 64 nodes, respectively, with sigmoid activation. After the first layer, a drop probability of 0.25 is applied, and after the second layer, a drop probability of 0.1 is used. The Adam optimizer is used during training, and the loss function is the mean squared error (MSE) function. The model was trained in 300 training epochs.

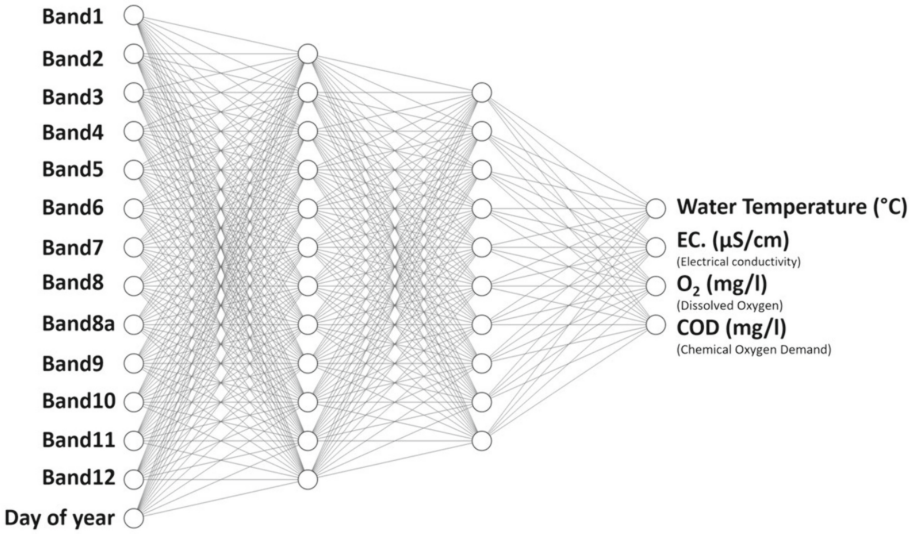

Fig. 5. DNN model used for prediction water quality indicators (T, EC, DO, and COD).

5 Results

The model was trained using the historical dataset regarding the HS Varadin Bridge. Verification of the accuracy of the predictions of the trained model is performed on the data collected at the HS Zemun. The dataset for checking the predictions is created similarly to the dataset for training the model. However, satellite data regarding the HS Zemun is downloaded only for dates when RHMZ and SEPA measurement data are available. A filter mask is applied to retain the values measured above the water surface (to remove the values measured above the land). Average values for each of the 13 bands are calculated from the remaining measurement values on water surfaces. In the continuation, the values are normalized to the range 0.1, and such values are divided by the network input. Finally, the network outputs predictions for four parameters (T, EC, DO, and COD) are obtained.

The results presented in Figs. 6 and 7 are predictions on data collected at HS Zemun, serving as a test dataset, since the network was trained using data collected at HS Varadin Bridge. The ML model prediction results demonstrate high accuracy in predicting T, EC, and DO from satellite data, with R2 values exceeding 0.98, indicating that the model explains over 98% of the variance in these parameters (Fig. 6). However, the prediction for COD is slightly less accurate, with an R2 value of 0.9209. The mean square errors (MSE) for the predicted values of the four parameters are 2.7845, 1269.82, 1.4778 and 1.6557. Figure 7 displays four line graphs showing the values for T, EC, and DO for the period from 2015 to 2023, where the values predicted by the trained model can be compared with the expected values.

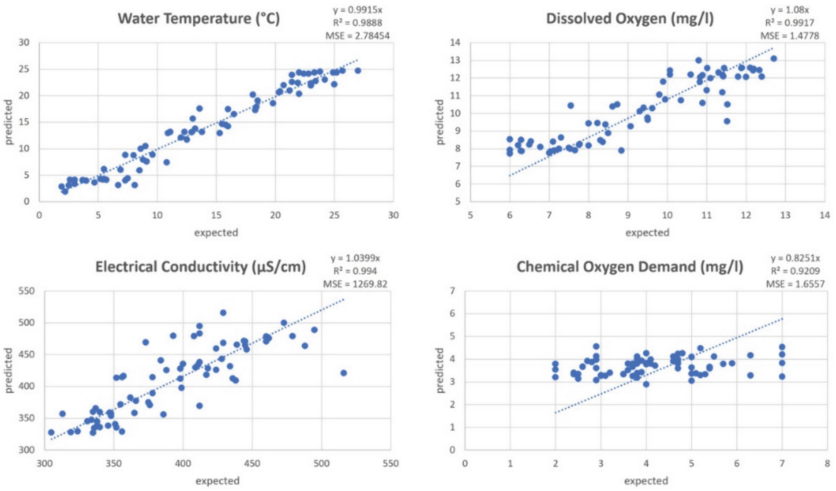

Fig. 6. Results of prediction at HS Zemun location.

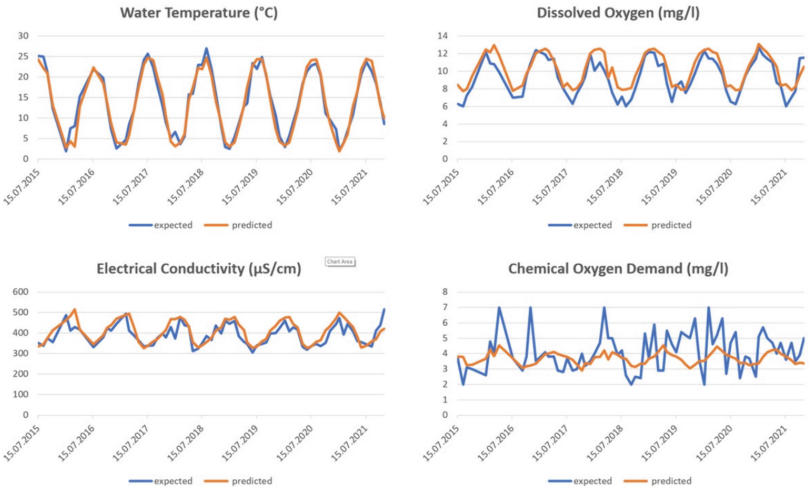

Fig. 7. Results of prediction at HS Zemun location on test results.

6 Conclusion

This study demonstrates that a ML model trained on Sentinel-2 satellite data is highly effective for predicting significant water quality parameters in the Danube River. By integrating multispectral satellite data with in-situ measurements, the model exhibits high accuracy in predicting significant water quality parameters, with an R2-result exceeding 0.98 for most indicators. Although slightly less precise for COD, the overall method marks a significant improvement over traditional techniques, combining the strengths of remote sensing and machine learning. This scalable approach not only enhances

monitoring at hydrology stations near urban areas along the Danube but also shows potential for application in other water bodies. The adaptability of our filtering mask technique further underscores its utility in diverse inland water surface contexts. In the next phase of our project, we plan to collect data from multiple measuring points along the Danube to develop a more comprehensive dataset. This expansion will facilitate more precise predictions across all parameters, with a particular focus on improving the accuracy of the COD parameter.

Acknowledgment. This paper results from project REmote WAter quality monitoRing anD IntelliGence – REWARDING [grant number 6707], supported by the Science Fund of the Republic of Serbia.

References

1. Gholizadeh, M.H., Melesse, A.M., Reddi, L.: A comprehensive review on water quality parameters estimation using remote sensing techniques. Sensors **16**(8), 1298 (2016). https://doi.org/10.3390/s16081298
2. Huangfu, K., Li, J., Zhang, X., Zhang, J., Cui, H., Sun, Q.: Remote estimation of water quality parameters of medium-and small-sized inland rivers using Sentinel-2 imagery. Water **12**(11), 3124 (2020). https://doi.org/10.3390/w12113124
3. Jiang, D., Wang, K.: The role of satellite-based remote sensing in improving simulated streamflow: a review. Water **11**(8), 1615 (2019). https://doi.org/10.3390/w11081615
4. Lioumbas, J., et al.: Satellite remote sensing to improve source water quality monitoring: a water utility's perspective. Remote Sens. Appl. Soc. Environ. **32**, 101042 (2023). https://doi.org/10.1016/j.rsase.2023.101042
5. Rahat, S.H., et al.: Remote sensing-enabled machine learning for river water quality modeling under multidimensional uncertainty. Sci. Total. Environ. **898**, 165504 (2023). https://doi.org/10.1016/j.scitotenv.2023.165504
6. Topp, S.N., Pavelsky, T.M., Jensen, D., Simard, M., Ross, M.R.: Research trends in the use of remote sensing for inland water quality science: moving towards multidisciplinary applications. Water **12**(1), 169 (2020). https://doi.org/10.3390/w12010169
7. Ritchie, J.C., Zimba, P.V., Everitt, J.H.: Remote sensing techniques to assess water quality. Photogramm. Eng. Remote Sens. **69**(6), 695–704 (2003). https://doi.org/10.14358/PERS.69.6.695
8. Guo, H., Huang, J.J., Chen, B., Guo, X., Singh, V.P.: A machine learning-based strategy for estimating non-optically active water quality parameters using Sentinel-2 imagery. Int. J. Remote Sens. **42**(5), 1841–1866 (2021). https://doi.org/10.1080/01431161.2020.1846222
9. Harkort, L., Duan, Z.: Estimation of dissolved organic carbon from inland waters at a large scale using satellite data and machine learning methods. Water Res. **229**, 119478 (2023). https://doi.org/10.1016/j.watres.2022.119478
10. Li, X., Ding, J., Ilyas, N.: Machine learning method for quick identification of water quality index (WQI) based on Sentinel-2 MSI data: Ebinur Lake case study. Water Supply **21**(3), 1291–1312 (2021). https://doi.org/10.2166/ws.2020.381
11. Tian, S., et al.: Remote sensing retrieval of inland water quality parameters using Sentinel-2 and multiple machine learning algorithms. Environ. Sci. Pollut. Res. **30**(7), 18617–18630 (2023). https://doi.org/10.1007/s11356-022-23431-9

12. Tadić, J.M., Ilić, V., Ilić, S., Pavlović, M., Tadić, V.: Hybrid machine learning and geostatistical methods for gap filling and predicting solar-induced fluorescence values. Remote Sens. **16**(10), 1707 (2024)
13. Tadić, J.M., et al.: Elliptic Cylinder Airborne Sampling and Geostatistical Mass Balance Approach for Quantifying Local Greenhouse Gas Emissions. ACS American Chemical Society, Environmental Science & Technology (2017)
14. Tadić, J.M., Ilić, V., Biraud, S.: Examination of geostatistical and machine-learning techniques as interpolators in anisotropic atmospheric environments. Atmos. Environ. **111**, 28–38, Elsevier Ltd. (2015). ISSN: 1352–2310
15. Dodig, A., Ricci, E., Kvascev, G., Stojkovic, M.: A novel machine learning-based framework for the water quality parameters prediction using hybrid long short-term memory and locally weighted scatterplot smoothing methods. J. Hydroinformatics, jh2024273 (2024)
16. Cojbasic, S., et al.: Application of machine learning in river water quality management: a review. Water Sci. Technol. **88**(9), 2297–2308 (2023)
17. RHMZ Homepage. https://www.hidmet.gov.rs/. Accessed 15 Oct 2023
18. SEPA Homepage. http://www.sepa.gov.rs/index.php. Accessed 15 Oct 2023

XGBoost "is All You Need": the Case of Forecasting Transmitted Heat Energy in District Heating Systems

Milan Zdravković(✉)

Faculty of Mechanical Engineering in Niš, University of Niš, Ul. Aleksandra Medvedeva 14, 18000 Niš, Serbia
milan.zdravkovic@masfak.ni.ac.rs

Abstract. This paper presents a comparative study of two distinct approaches, XGBoost and Long-Short Term Memory (LSTM), for forecasting transmitted heat energy in District Heating Systems (DHS). The objective is to explore scenarios in which conventional ML algorithms demonstrate better performance over deep learning networks in time series forecasting and the associated benefits in terms of computational cost and environmental impact. The study focuses on a real-world DHS dataset. Through experimentation and analysis, it is demonstrated that XGBoost consistently outperforms LSTM in this specific forecasting task. The difference is explained by the error distribution illustrating that LSTM makes more significant errors in the intervals of less data availability. The reduced computational demands of conventional ML approaches not only result in cost savings but also minimize the carbon footprint associated with data analysis tasks in energy systems.

Keywords: Machine Learning · Neural Networks · Time Series Forecasting · District Heating Systems

1 Introduction

The extreme popularity and availability of off-the-shelf deep learning algorithms and architectures today have created excitement and very high expectations related to quickly addressing different automation challenges in different industries. The promise of simplicity of use combined with performance already demonstrated mostly in the cases of language processing and computer vision, has led to a surge in applying these technologies across various sectors. These include healthcare, finance, automotive, and more, where they are being used for tasks like disease detection, financial forecasting, autonomous driving, and customer service automation. However, this enthusiasm must be tempered with a recognition of the complexities involved. Successful implementation often requires large amounts of high-quality data, and the ability to interpret and fine-tune models to specific needs. Moreover, issues like algorithmic bias, transparency, and ethical considerations pose additional challenges.

M. Trajanović et al. (Eds.): ICIST 2024, LNNS 860, pp. 12–22, 2024.
https://doi.org/10.1007/978-3-031-71419-1_2

The reality is that while deep learning offers powerful tools, their effective application demands more data, more care and more expertise. Quite often, Deep Learning architectures are tested quickly and applied without careful consideration and with prejudice driven by AI hype. This hasty adoption often leads to overlooking crucial aspects like algorithm suitability, data quality, and computational requirements. The result is systems that either underperform or consume excessive energy, thus negating the benefits of using AI. This situation underscores the importance of a more measured approach to implementing Machine Learning solutions, one that involves thorough testing, consideration of environmental impact, and an understanding of the specific problem context. Moreover, it emphasizes the need for organizations to invest in building or acquiring the necessary expertise to harness the full potential of AI technologies effectively and sustainably.

The objective of research behind this paper is to demonstrate that traditional ML algorithms are indeed competitive when compared to complex neural network algorithms in certain time series forecasting problems. This will be showcased on the example of forecasting transmitted heat energy in District Heating Systems.

Despite the maturity of District Heating systems (DHS), substantial opportunities exist for enhancing their operational efficiency. This particularly pertains to the reduction of fuel consumption costs and minimization of carbon emissions. A promising strategy involves capitalizing on the considerable potential for reengineering current short- and long-term operational strategies of DHS. This can be achieved through the utilization of precise heat demand forecasts. Such forecasts are crucial for optimizing heat production, which leads to reduced fuel consumption, waste, and CO2 emissions. Additionally, this approach ensures maximum consumer satisfaction and facilitates more effective planning for both short and long-term heat production.

The forecasting problem and methodology are described in Sect. 2. Section 3 introduces and describes two competitive approaches, and it presents the implementation of experiments with the two proposed and argued architectures. Section 4 discusses the final results.

2 Methodology

In essence, the operation of District Heating System (DHS) plants involves either automated or semi-automated management of primary (at the plant level) and secondary (at the substations level) supply water temperatures, as well as water flow in the primary supply. The primary and secondary flows are closed loop and the energy from primary to secondary lines is exchanged through a heat exchanger. The management of supply water temperature is based on the collective DHS demand and prevailing weather conditions. The overall DHS demand is determined by the transmitted heat energy in the specified interval, namely the difference between the measurements of transmitted energy in the current and past timepoint, recorded at the calorimeter, located at the return primary line.

Presently, conventional DHSs are managed through a Supervisory Control And Data Acquisition (SCADA) system. This system integrates various sensors, control mechanisms, and algorithms that automatically modify operational parameters in response to sensor data. DHS control at the district heating substation levels (primary and secondary

sides of DHS) is automated. This includes the implementation of appropriate hot water reset controls (outdoor air reset or control curve), often described as a regulation curve, used by SCADA system to deduce desired supply line water temperature based on the measured outside air temperature.

The process is described in detail and forecasting model elaboration is provided in the earlier work [1]. This paper examines the potential to replace simplistic control curve with a model capable to forecast the transmitted energy based on the measured air temperature in the previous timepoint.

The selected comparative methods are stacked Long-Short Term Memory architecture and Gradient Boosting approach, namely its XGBoost implementation. The experiment involves preparing the data appropriately for each model, training both models, and then evaluating and comparing their performance using relevant metrics. Additionally, XGBoost model is optimized by finding the set of hyperparameters providing the best metrics. The method used was Bayesian optimization.

The metrics used were Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Coefficient of determination (R2 score). The coefficient of determination, often denoted as R^2 (R-squared) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Additional metrics for comparison are time for training and inference on the test set. The experiment is carried out in Google Colab environment, using T4 GPU runtime.

Implementation of XGBoost for regression problems expects a structured dataset, not a time series. Transforming time series data into a format suitable for regression problems is a common approach in Machine Learning for forecasting and other time-dependent analyses. This process involves converting the sequential nature of time series data into a structured format that a regression model can understand. One standard method is to create lagged features, which are values from previous time steps used as separate input features. The number of lagged features (also known as the lag order) depends on the specific problem and how far back in the past the predictive patterns extend. Additionally, certain qualities of the times of the measured instances will be extracted and used as features, namely, hour of the day, day of the week and month.

3 Implementation

Data from one substation, namely substation 9 from the local DHS, was used for the experiment. Data included outside ambient temperature from the sensor located at the building facade, on the north side; temperatures of water in supply and return primary and secondary lines and transmitted energy, measured by the calorimeter located at the primary return line. Two heating seasons were considered for analysis, namely 2018/19 and 2019/20, in total 5832 timepoints.

Ambient temperature and transmitted energy were used in model training, while the other features were dropped. Transmitted energy in the selected timepoint is calculated as the difference of the energy reading in that timepoint and in the previous one. Only the period from November to March was considered in the analysis. Normally, heating season starts mid-October and lasts till mid-April. However, those periods are characterized by high variance in temperature, special regimes of operation and thus, will not

be accounted for. The missing data was found in the dataset. This was due to the lack of 3G network connectivity at certain time points. Missing data were imputed by using linear interpolation.

Overall distribution of available data is displayed in Fig. 1, after resetting date time index in order to enable a continuous signal presentation. Also, zoomed in overlay plot illustrates the correlation between the outside temperature and transmitted heat energy.

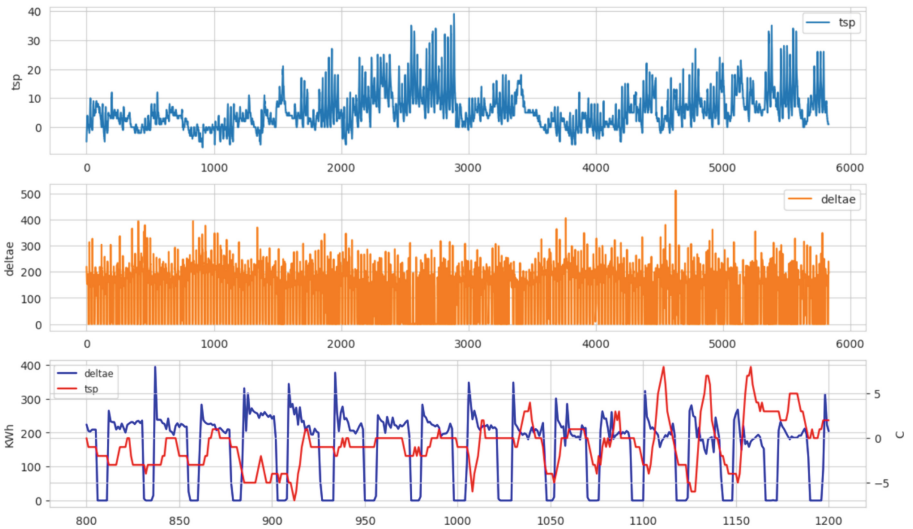

Fig. 1. Ambient temperature (top), transmitted energy time series data (mid) and zoomed-in overlay of two signals.

Both time series are stationary, which is confirmed by the Augmented Dickey-Fuller (ADF) test. In a stationary time series, the mean, variance, and autocorrelation structure remain constant across different time points. Stationarity is an important requirement for the good performance of parametric algorithms, such as neural networks.

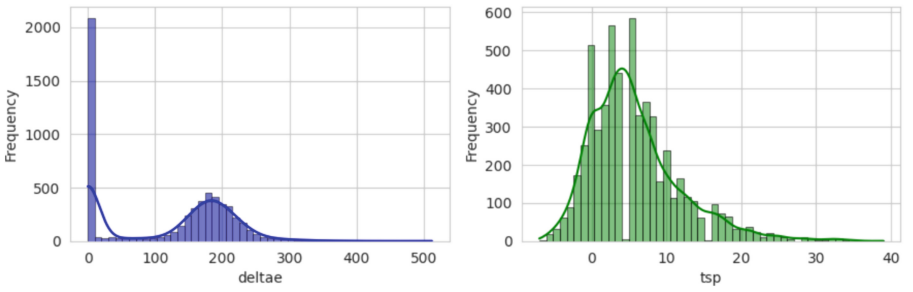

Fig. 2. Distribution of transmitted energy (left) and ambient temperature (right) with KDE line.

The distribution analysis of two relevant time series signals (distribution of data points - histogram, and the underlying probability density - Kernel Density Estimate plot are presented in a Fig. 2) indicate relatively high sparsity of hourly transmitted energy feature, where zeros in certain time points indicate that the system is not operational, mostly in cases of high outside temperatures. When this is ignored, deltae signal exhibits normal distribution. Normal distribution is also exhibited by the outside temperature signal, with minor skewness towards higher temperatures.

The value of Spearman coefficient (-0.308) and p-value (0.000) indicate that there is statistically significant negative association between two signals, which is expected. Even though both signals are normally distributed, Spearman values are considered instead of Pearson coefficients as more reliable indicator of association because of relatively high sparsity of deltae signal and zero data which can be also interpreted as outliers to which Spearman coefficient exhibits better response. Besides less sensitivity to outliers, the Spearman coefficient indicates monotonic relationship and does not assume linear association.

Two models and their respective forecasting capabilities will be tested with the data above, namely stacked LSTM model and XGBoost.

3.1 Implementation of Stacked LSTM Model

Long Short-Term Memory (LSTM) networks [2] are a special kind of Recurrent Neural Network (RNN) [3], specifically designed to learn from sequences of data and remember long-term dependencies in the data. They are widely used for sequence prediction problems, such as time series forecasting, natural language processing, and speech recognition. LSTMs are design to address the specific limitation of traditional RNNs related to struggling to capture long-term dependencies in a sequence due to the vanishing gradient problem.

LSTMs maintain a hidden state vector and a cell state vector across time steps, which help them store and manage long-term dependencies in the data. An LSTM unit consists of three main components: forget gate which “decides” what information should be thrown away or kept, input gate that updates the cell state, and output gate which “decides” what the next hidden state should be. In each time step, the LSTM cell takes three pieces of information: the current input data, the previous hidden state, and the previous cell state. Based on these inputs, it produces a new hidden state and a new cell state, which are passed to the next time step.

LSTM networks take structured data transposed to supervised regression problem format by introducing certain number of datapoints from the past in one instance, namely lagged timepoints. For both experiments, 6 timepoints in the past will be considered in each data instance. 80% of all data will be used for training, while 20% will be set aside for testing the trained model.

For the case of training LSTM architecture, data was normalized. Normalization is a crucial step for pre-processing data for training neural networks, as it improves gradient descent efficiency and effectiveness by preventing local minima.

The architecture used in the experiment was stacked LSTM, with two LSTM layers each with 100 units and Rectified Linear Unit activation, each followed with dropout layer. Mean absolute error was used as a loss function and efficient Adam optimizer has

been used. Training was carried out with 100 epochs and batch size of 24, where 20% of the training set was used for validation.

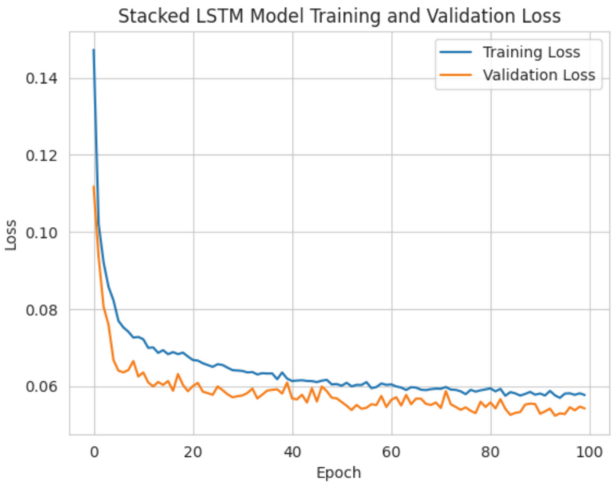

Fig. 3. Stacked LSTM model training and validation loss

Model training and validation loss curve (see Fig. 3) demonstrated good generalization and it did not show overfitting, already mitigated by using dropout regularization in the model architecture.

3.2 Implementation of XGBoost Model

XGBoost (Extreme Gradient Boosting) [5] is a highly efficient and flexible algorithm widely used for supervised learning tasks. It is an implementation of gradient boosted decision trees designed for speed and performance. XGBoost has gained popularity in machine learning competitions and practical applications due to its effectiveness and efficiency in handling various types of tabular data and relevant tasks.

XGBoost is an ensemble learning method, specifically a boosting technique. It builds the model in stages, and each stage adds new models to correct the errors made by the existing ensemble of models. It operates within the gradient boosting framework [6] by constructing a new model that adds to an existing ensemble of models in a way that minimizes the overall prediction error. The “gradient boosting” part refers to the algorithm’s use of the gradient descent algorithm to minimize the loss when adding new models. XGBoost primarily uses decision trees as its base learners. Each new tree corrects the residual errors (differences between predicted and actual values) of the previous trees. A key feature that differentiates XGBoost from other gradient boosting methods is its built-in regularization (both L1 and L2), which helps to prevent overfitting and improve model generalization. XGBoost can automatically handle missing data, making it robust to problems with incomplete datasets. It is optimized to efficiently handle sparse data. It incorporates a learning rate (shrinkage), which scales the contribution of each new tree added to the model. This can be used to prevent overfitting.

Feature engineering practices ensure that time dimensions are accurately reflected in the relevant features. Given the cyclic nature of district heating systems operation, it is assumed that hour of day is one of the significant features. Besides that, relevance of day of the week is expected to be non-trivial, especially when considering if the day is a working day or not. Spearman correlation coefficients show statistically significant association between hour of the day and month with transmitted heat energy ($SP_{hour} = -0.339$, $SP_{mon} = -0.150$). Interestingly, the hypothesis on the association between day of the week and transmitted heat energy was not confirmed.

For optimizing hyperparameters of XGBoost regressor, a Bayesian approach was used. Bayesian optimization is an optimization strategy, particularly useful for Machine Learning scenarios where the evaluation of the objective function (such as model validation loss) is computationally expensive. Bayesian optimization is a probabilistic model-based approach. It constructs a posterior distribution of functions (probability model) that best describes the function that needs optimization, based on past evaluations. The process is iterative, and it starts with a set of initial hyperparameter combinations (often chosen randomly). Then, modeling the objective function is carried out, by using the results from initial and ongoing evaluations.

In the optimization case, MAE was used as the objective function. The space of hyperparameters was defined. For implementation of the Bayesian optimization, Hyperopt [7] package was used. It is an open-source Python library used for optimizing the hyperparameters of machine learning algorithms. The Tree of Parzen Estimators (TPE) was used as an optimization algorithm. The optimization process was carried out in 80 iterations.

4 Discussion of Results

XGBoost model outperformed LSTM model in all metrics (see summary results in Table 1). Optimization somewhat improved the performance of the model with default set of hyperparameters. In both cases, RMSE was almost double the Mean Absolute Error (MAE). RMSE gives more weight to larger errors due to the squaring of each error before averaging, while MAE treats all errors equally. RMSE being significantly higher than MAE suggests wider spread of errors or the presence of some large errors in predictions. In general, the model shows good accuracy, but it makes a few substantial errors.

The histograms displayed in Fig. 4 show the distribution of errors (the differences between predicted and actual values) in the case of optimized XGBoost and LSTM approach. The red dashed line at zero visualize the point where there is no error (perfect prediction). The Kernel Density Estimate (KDE) line provides a smooth curve representing the error density.

Error distribution histogram confirmed the assumption on good accuracy with few substantial errors in using both approaches made earlier. However, it also explained the difference in performance of both methods. Occurrence of minor fat tails - increased accumulation of larger errors in case of LSTM models explains the source of difference in performance metrics: while LSTM model has more successful forecasts with less error, it also makes more substantial errors that affect the MAE and especially RMSE.

Table 1. Summary results of experiment

	LSTM	XGB	XGB Optimized
RMSE (kW)	62.111	36.700	35.677
MAE (kW)	28.890	20.589	18.687
R2	0.540	0.853	0.861
Training time (sec)	476.129	0.238	549.237
Inference time (sec)	0.537	0.007	0.035
Training CO2 (gr)	9.245	0.005	10.664
Inference time (per 1000, gr)	10.421	0.141	0.676

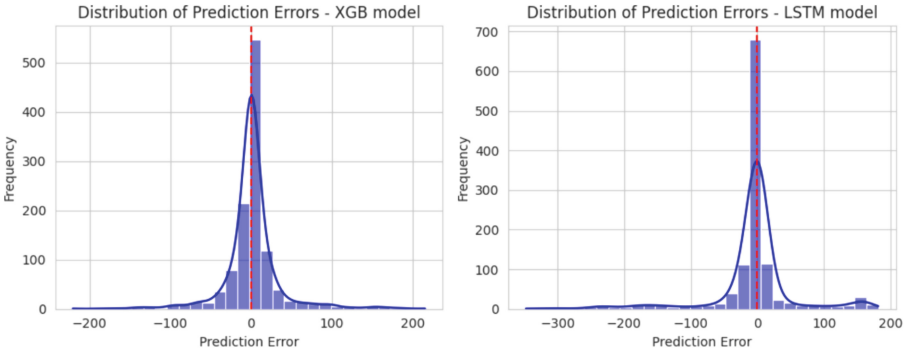

Fig. 4. Distributions of prediction errors by XGBoost (left) and stacked LSTM model (right)

Figure 5 shows the scatter plot of actual and predicted values for the optimized XGBoost model and LSTM model, as well as overlayed distribution of test data.

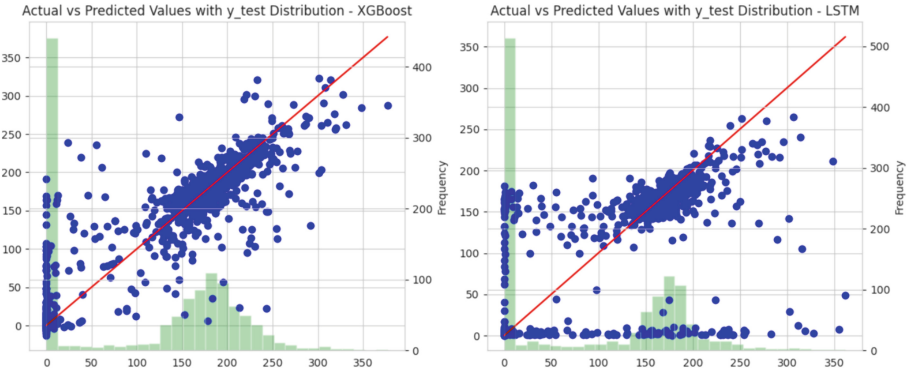

Fig. 5. Plot of actual and predicted values with overlayed distribution of test data

The figure illustrates the accuracy of the model in the different intervals of test data. In both cases, the scatter plot points are clustered along the red line, which suggests the predictions are reasonably good. However, there is some variance in both models (more significant in LSTM than XGBoost), especially for lower and higher values where the points tend to diverge more from the line. This variance can be explained by lack of data in intervals of lower and higher values for transmitted energy.

Additionally, scatter plots unveil the specific nature of significant errors made by both models, mostly due to impossibility to forecast the periods in which the heating is turned off by the operator (for XGBoost and LSTM models, concentration of forecasts along Y-axis) or to confusing those periods with periods in which the heating is actually on (LSTM models, concentration of forecasts along X-axis). It's important to emphasize that the decision to turn on the heating system lies with a human operator, and it is affected by the reasons not included as features in this dataset.

In general, the LSTM models are expected to show good performance at uncovering very complex patterns of heat demand that occur at the beginning and end of the heating season, when heat demand, as well as outside temperature exhibit high variance. However, there is no sufficient data to unleash the power of LSTM's long-term memory in the case of this experiment.

Furthermore, neural networks require data imputation. In this case, occasional missing data is replaced by using simplistic linear interpolation technique. It's worth highlighting that many of the data imputation approaches (besides stochastic ones) introduce regularities that are considered as bias that can lead to better results than in reality. XGBoost assumes initial transformation of sequential time series data to structured, tabular data suitable for traditional regression problems. This appears very useful in industrial application where the periods of missing data due to sensor faults are frequent. Such faults and corresponding missing data problems cannot be addressed with imputation techniques, since those periods can be quite long. Ignoring sequential nature of data dramatically improves usability of data islands, occurring in such circumstances.

What are the possible reasons for XGBoost outperforming LSTM architecture in this case? First of the reasons is the size of the dataset combined with the time feature engineering practices. XGBoost can benefit significantly from good feature engineering if those features encapsulate the temporal dynamics well. LSTM is indeed very good in uncovering these dynamics, but only if there is sufficient data available. Second reason is that LSTMs can be particularly sensitive to the choice of hyperparameters, especially when considering the network topology. The optimization has not been done in this case as it would require significant computational resources and time.

Besides accuracy, computational requirements for the two methods are not comparable, especially when inference is considered. Training times in both cases appear to be similar, but only when optimization of hyperparameters is involved in XGBoost case. The amount of energy used for training a neural network on a T4 GPU depends on several factors, including network architecture complexity, training dataset size, batch size, choice of optimizer and learning rate. Based on the technical specifications of the manufacturer[1] the power usage of NVidia T4 GPU is 70 watts, corresponding to the

[1] https://www.nvidia.com/en-eu/data-center/tesla-t4/.