Prediction of Jakarta's Air Quality Using a Stacking Framework of CLSTM, CatBoost, SVR, and XGBoost

Usman Syapotro[1], Silvia Ratna[2], M. Muflih[3], Haldi Budiman[4], M. Rezqy Noor Ridha[5], Muhammad Hamdani[6]

[1,2,3,4,5,6]Faculty of Information Technology, Islamic University of Kalimantan Muhammad Arsyad Al-Banjari, Indonesia

Email: 05usman.syapotro.study@gmail.com[1], silvia.ratna@uniska-bjm.ac.id[2], mmuflihfti@uniska-bjm.ac.id[3], haldibudiman@uniska-bjm.ac.id[4], rezqyridha@gmail.com[5], mhdhamdani.formal@gmail.com[6]

Abstract

Air quality prediction, particularly in estimating PM10 particle concentration, is a significant challenge in major cities like Jakarta, which experience high levels of air pollution. This study aims to develop an air quality prediction model using an innovative stacking framework that combines several machine learning algorithms, namely ConvLSTM, CatBoost, SVR, and XGBoost. The methodology employed in this research is an experimental approach, where each model is trained and tested individually before being integrated into the stacking framework. The dataset used was sourced from the Kaggle platform, containing historical air quality data in Jakarta. Performance evaluation was conducted by measuring the Root Mean Squared Error (RMSE) for each model. The results of the study showed that the ConvLSTM model produced an RMSE of 13.5168, CatBoost with an RMSE of 13.4113, and SVR with an RMSE of 14.2725. To improve prediction accuracy, the researchers employed a stacking approach of the four models (ConvLSTM, CatBoost, SVR, and XGBoost), which yielded a much lower RMSE of 0.8093. Thus, this stacking framework has proven to significantly enhance air quality prediction performance, particularly in predicting PM10 concentrations in Jakarta.

Keywords

Prediction of Jakarta Air Quality, PM 10, Stacking, Extreme Gradient Boosting (XGBoost)

Introduction

Air quality is a significant issue in many large cities, particularly in densely populated urban areas like Jakarta. Air pollution caused by motor vehicle emissions, industrial activities, and the burning of fossil fuels has become a serious threat to public health and the environment. Recent studies show that long-term exposure to air pollution, particularly PM10 particles, can accelerate the development of coronary heart disease, adding to the existing healthcare burden (Urbanowicz et al., 2024). Additionally, research also indicates that high exposure to PM10 may increase the risk of autoimmune diseases, especially rheumatoid arthritis, further reinforcing the evidence of the harmful impact of air pollution on human health (Adami et al., 2022)

Submission: 20 October 2024; **Acceptance:** 21 November 2024

(Handhayani, 2023) investigated the relationship between air pollution and meteorological conditions in Jakarta, utilizing a column-based data integration model and the PC algorithm to establish causal relationships. The study further employed Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models to forecast the Air Quality Index (AQI) and meteorological conditions, demonstrating the superior performance of LSTM when using integrated data. However, the dynamic and complex nature of air pollution necessitates the exploration of more sophisticated techniques, such as the hybridization of machine learning models, to improve predictive capabilities. A study by (Yunis et al., 2024) exemplifies this approach, demonstrating the effectiveness of combining models like SVR, SARIMA, LSTM, and Prophet to achieve higher accuracy in forecasting air pollution levels. The integration of meteorological data and the utilization of hybrid models offer a promising avenue for developing robust air quality prediction systems, ultimately contributing to informed decision-making and effective pollution mitigation strategies.

While previous studies have explored air quality prediction using various models and approaches, this research offers novelty by combining the strengths of Convolutional Long Short-Term Memory (CLSTM), CatBoost, Support Vector Regression (SVR), and XGBoost within a previously unexplored stacking framework for Jakarta air quality prediction. This framework aims to overcome the limitations of each model individually and leverage their combined advantages to achieve more accurate and reliable predictions of air quality in Jakarta.

The remainder of this paper consists of methodology, results and discussion, and conclusions and recommendations. The methodology section discusses the research methodology employed, including data collection methods, data processing, and framework development. The results and discussion section presents the experimental results, which are analyzed in-depth to gain a comprehensive understanding. The final section, conclusions and recommendations, presents the conclusions of this research along with suggestions for future research.

Methodology

This research employs a time series prediction approach using several machine learning algorithms, namely ConvLSTM, CatBoost, SVR, and a stacking model with XGBoost as the meta-learner. The following are the methodological steps:

Data Collection and Preparation

Time series data is processed by converting the date column to 'YYYY-MM' format. The data is then split into 80% for training and 20% for testing.

Dataset Formation with Timesteps

A dataset is shaped using a 24-month time window to predict the subsequent month. A dedicated function is used to generate feature (X) and target (y) pairs based on this time window.

ConvLSTM Model

The ConvLSTM model is designed to capture temporal patterns in time series data (Esquivel et al., 2020) The data is reshaped into a 5D format to align with the model's architecture and trained to predict the target value.

CatBoost Model

A CatBoost model is applied to handle time series prediction without requiring complex data reshaping, as demonstrated in previous research where CatBoost was successfully employed for similar time series forecasting problems (Prokhorenkova et al., 2018).

SVR Model

SVR (Support Vector Regression) is a popular machine learning algorithm widely used in regression tasks. It transforms the nonlinear relationship between the input vector and the corresponding real response into a linear relationship in a higher-dimensional feature space through an unknown function (Borrero & Mariscal, 2023). SVR's flexibility and ability to handle nonlinear data make it an attractive choice for time series prediction.

Stacking Model with XGBoost

Prediction results from ConvLSTM, CatBoost, and SVR are combined as input for the stacking model, which employs XGBoost as the meta-learner. In related work, Luo et al. (2024) proposed a novel algorithm for short-term load forecasting. Their approach also leverages a Stacking ensemble algorithm, combining Convolutional Neural Network-Bidirectional Long Short-Term Neural Network-Attention Mechanism (CNN-BiLSTM-Attention) with Extreme Gradient Boosting (XGBoost).

Model Evaluation

Each model is evaluated using the Root Mean Squared Error (RMSE) metric, with a focus on comparing the performance between individual models and the stacking model.

Experimental Results

The stacking model is expected to deliver more accurate results compared to the individual models by combining the strengths of each base model.

Results and Discussion

Based on the calculation of the Root Mean Squared Error (RMSE) from various machine learning models used in this research, the following values were obtained:

Table 1 Model Accuracy

Model	Dataset Split	RMSE
ConvLSTM	80% Train, 20% Test	13.5168
CatBoost	80% Train, 20% Test	13.4113
SVR	80% Train, 20% Test	14.2725
Framework Stacking (XGBoost)	**80% Train, 20% Test**	**0.8093**

From Table 1, it is evident that the stacking model, which combines the predictions of the ConvLSTM, CatBoost, and SVR models with XGBoost as the meta-learner, yields significantly better results compared to each individual model.

Research focuses on predicting air quality in Jakarta, specifically estimating the concentration of PM10 particles. Given the detrimental effects of air pollution on public health and the environment, this issue is of paramount importance. As a major city with high levels of air pollution, Jakarta is the primary focus of this study. The objective is to develop an accurate air quality prediction model by utilizing an innovative stacking framework that combines several machine learning algorithms, namely ConvLSTM, CatBoost, SVR, and XGBoost.

Research methodology employs an experimental approach. Each model is trained and tested individually before being integrated into the stacking framework. The dataset, sourced from the Kaggle platform, contains historical air quality data for Jakarta. Model performance is evaluated using the Root Mean Squared Error (RMSE) metric. Research findings indicate that the ConvLSTM model yields an RMSE of 13.5168, CatBoost an RMSE of 13.4113, and SVR an RMSE of 14.2725.

To enhance prediction accuracy, a stacking approach combining four models (ConvLSTM, CatBoost, SVR, and XGBoost) is implemented, resulting in a significantly lower RMSE of 0.8093. This result demonstrates that the stacking framework significantly improves air quality prediction performance, particularly in predicting PM10 concentrations in Jakarta.

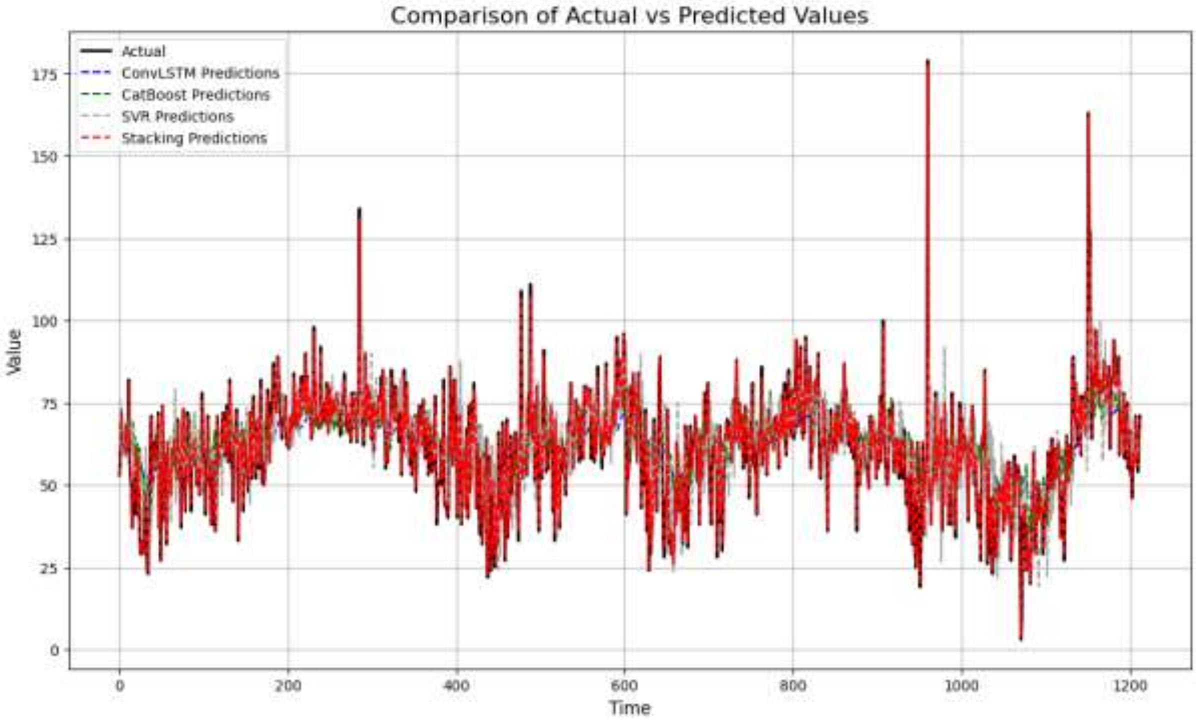

Figure 1 Prediction Results Graph

Description:

Black	: Actual PM 10 values
Blue	: ConvLSTM model predictions
Green	: CatBoost model predictions

Darkgray	: SVR model predictions
Red	: Framework Stacking (XGBoost)

Figure 1, titled "Prediction Results Graph," presents a comparison of PM 10 predictions from four machine learning models ConvLSTM, CatBoost, SVR, and a Stacking framework (XGBoost) against actual values over time. Overall, the Stacking framework (XGBoost) delivers the closest predictions to the actual PM 10 values, as evidenced by the red line closely tracking the pattern of the black line. While other models, such as CatBoost also shows good performance, Stacking outperforms these base models, particularly in predicting sharper fluctuations. This highlights the Stacking framework's ability to leverage the strengths of multiple individual models to generate more accurate and stable predictions.

Conclusion

This study examined the crucial problem of predicting Jakarta's air quality, with an emphasis on measuring PM10 particle concentrations. To improve prediction accuracy and dependability, the study used a novel stacking framework that included the benefits of ConvLSTM, CatBoost, SVR, and XGBoost. With an extremely low RMSE of 0.8093 when compared to the separate models, the results clearly showed the superiority of the stacking model. This demonstrates how well the stacking strategy works to combine the complimentary strengths of many machine learning algorithms to get predictions that are more accurate and reliable. The research has significant significance and provides a useful tool for public health organizations, environmental agencies, and lawmakers to make judgments and put into practice efficient plans to lessen the negative consequences.

Recommendation

Future research directions include leveraging larger datasets to enhance the accuracy and generalization of the prediction model, as well as exploring a wider range of stacking configurations, incorporating diverse combinations of algorithms and prediction aggregation methods, to identify the optimal configuration for achieving the best performance.

References

Adami, G., Pontalti, M., Cattani, G., Rossini, M., Viapiana, O., Orsolini, G., Benini, C., Bertoldo, E., Fracassi, E., Gatti, D., & Fassio, A. (2022). Association between long-term exposure to air pollution and immune-mediated diseases: A population-based cohort study. *RMD Open*, *8*(1), 1–8. https://doi.org/10.1136/rmdopen-2021-002055

Borrero, J. D., & Mariscal, J. (2023). Elevating Univariate Time Series Forecasting: Innovative SVR-Empowered Nonlinear Autoregressive Neural Networks. *Algorithms*, *16*(9), 1–15. https://doi.org/10.3390/a16090423

Esquivel, N., Nicolis, O., Peralta, B., & Mateu, J. (2020). Spatio-Temporal Prediction of Baltimore Crime Events Using CLSTM Neural Networks. *IEEE Access*, *8*, 209101–209112. https://doi.org/10.1109/ACCESS.2020.3036715

Handhayani, T. (2023). An integrated analysis of air pollution and meteorological conditions in Jakarta. *Scientific Reports*, *13*(1), 1–11. https://doi.org/10.1038/s41598-023-32817-9

Luo, S., Wang, B., Gao, Q., Wang, Y., & Pang, X. (2024). Stacking integration algorithm based on CNN-BiLSTM-Attention with XGBoost for short-term electricity load forecasting. *Energy Reports*, *12*(May), 2676–2689. https://doi.org/10.1016/j.egyr.2024.08.078

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *2018-December*(Section 4), 6638–6648. https://dl.acm.org/doi/abs/10.5555/3327757.3327770

Urbanowicz, T., Skotak, K., Olasińska-Wiśniewska, A., Filipiak, K. J., Bratkowski, J., Wyrwa, M., Sikora, J., Tyburski, P., Krasińska, B., Krasiński, Z., Tykarski, A., & Jemielity, M. (2024). Long-Term Exposure to PM10 Air Pollution Exaggerates Progression of Coronary Artery Disease. *Atmosphere*, *15*(2), 1–13. https://doi.org/10.3390/atmos15020216

Yunis, R., Andri, A., & Djoni, D. (2024). Hybridization Model for Air Pollution Prediction Using Time Series Data. *CogITo Smart Journal*, *10*(1), 422–435. https://doi.org/10.31154/cogito.v10i1.619.422-435