Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 8 Nomor 1, Maret 2024, pp. 255-262
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Forecasting Climate Change Impacts Using Machine Learning and Deep Learning: A Comparative Analysis

Gregorius Airlangga
Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia
e-mail: gregorius.airlangga@atmajaya.ac.id

Abstract

This study undertakes a comparative analysis of machine learning and deep learning models for forecasting the impacts of climate change, utilizing Cross-Validation Root Mean Squared Error (CV RMSE) to gauge performance. Analyzed models include Long Short-Term Memory (LSTM) networks (CV RMSE: 0.155), Linear Regression (CV RMSE: 5647815244.91), Random Forest (CV RMSE: 0.159), Gradient Boosting Machine (GBM) (CV RMSE: 0.164), Support Vector Regressor (SVR) (CV RMSE: 0.159), Decision Tree Regressor (CV RMSE: 0.199), and K-Nearest Neighbors (KNN) Regressor (CV RMSE: 0.166). The study rigorously processes climate change time series data to ensure accurate, generalizable results. LSTM networks demonstrated exceptional performance, indicating their strong capacity for modeling complex temporal sequences inherent in climate data, while Linear Regression lagged significantly behind, revealing limitations in addressing non-linear patterns of climate change. The promising results of Random Forest and SVR models suggest their applicability in environmental science forecasting tasks. Our findings offer valuable insights into the efficacy of various predictive models, aiding researchers and policymakers in leveraging advanced analytics for climate change mitigation strategies.

***Keywords**: Climate Change Forecasting, Machine Learning, Deep Learning, Time Series Analysis, Predictive Models.*

Abstrak

Studi komprehensif ini menyelidiki secara komparatif berbagai metode pembelajaran mesin dan deep learning dalam konteks prediksi dampak perubahan iklim, mengadopsi Cross-Validation Root Mean Squared Error (CV RMSE) sebagai metrik evaluasi. Dalam analisis ini, berbagai model seperti jaringan Long Short-Term Memory (LSTM) (CV RMSE: 0.155), Regresi Linier (CV RMSE: 5647815244.91), Random Forest (CV RMSE: 0.159), Gradient Boosting Machine (GBM) (CV RMSE: 0.164), Support Vector Regressor (SVR) (CV RMSE: 0.159), Regressor Pohon Keputusan (CV RMSE: 0.199), dan Regressor K-Nearest Neighbors (KNN) (CV RMSE: 0.166) dianalisis secara mendalam. Penelitian ini menggarap data seri waktu perubahan iklim dengan teliti, menjamin integritas dan aplikabilitas temuan yang dihasilkan. Jaringan LSTM mencatat performa istimewa, menegaskan kapabilitas unggulnya dalam menggambarkan dependensi temporal kompleks yang terkandung dalam data iklim, berbanding terbalik dengan Regresi Linier yang menunjukkan keterbatasan signifikan dalam menangkap pola non-linear yang sering muncul dalam fenomena perubahan iklim. Sementara itu, kinerja mengesankan dari model Random Forest dan SVR menyarankan keefektifan mereka dalam aplikasi peramalan lingkungan. Hasil studi ini memberikan perspektif berharga mengenai kinerja relatif dari beragam model prediktif, mendukung para peneliti dan pembuat kebijakan dalam mengadopsi teknik analisis data canggih untuk formulasi strategi mitigasi perubahan iklim.

***Kata kunci**: Peramalan Perubahan Iklim, Pembelajaran Mesin, Deep Learning, Analisis Seri Waktu, Model Prediktif.*

1. INTRODUCTION

In In the realm of data science and machine learning, forecasting time series data related to climate change represents a critical and complex challenge [1]–[3].

This challenge spans across multiple domains, emphasizing the urgent need to accurately predict climate phenomena to address environmental concerns and inform policy-making [4]–[6]. The unique characteristics of climate data, including its temporal dependencies, non-stationarity, and potential for high dimensionality, demand the use of sophisticated analytical methods capable of effectively capturing these intricate dynamics [7]–[9]. The quest to develop robust models for forecasting climate change is motivated by the necessity for precise predictions that can support strategic environmental planning and promote sustainable practices [10]–[12].

A thorough examination of existing literature unveils a diverse range of methodologies that have been applied to tackle the intricacies of time series forecasting in the context of climate change [13]–[15]. While traditional statistical models like ARIMA (Autoregressive Integrated Moving Average) and exponential smoothing have provided a foundational understanding of temporal data analysis, the evolution of machine learning and deep learning technologies has considerably broadened the scope of predictive capabilities [16]–[18]. This advancement has introduced a variety of models adept at deciphering non-linear relationships and complex interactions characteristic of climate data [19]. Prominently, techniques such as Linear Regression, Random Forest, and Gradient Boosting Machines (GBMs) have been recognized for their adaptability and effectiveness in various forecasting applications [20]. Additionally, neural network-based methodologies, particularly Long Short-Term Memory (LSTM) networks, have proven to be invaluable for their ability to grasp long-term dependencies present in sequential climate data [21]. This integration of advanced computational models marks a significant step forward in the effort to enhance the accuracy and reliability of climate change forecasts, ultimately contributing to the global initiative of mitigating environmental risks and fostering resilience against climate variability and change [22].

Despite the strides made in developing effective forecasting models, the field is marked by an ongoing quest to address several pressing challenges [23]. One such challenge is the need for models that can adapt to the evolving statistical properties of time series data, a phenomenon commonly referred to as concept drift [24]. Additionally, the high dimensionality of some time series datasets poses significant challenges, necessitating models that can perform feature selection or dimensionality reduction implicitly to enhance prediction accuracy without compromising computational efficiency [25].

The goal of our research is to contribute to this evolving landscape by conducting a comprehensive evaluation of both traditional machine learning and advanced neural network-based models in the context of time series forecasting. Specifically, we aim to bridge the gap identified in the literature concerning the comparative analysis of these models under a unified framework [26]. This involves assessing the performance of a diverse set of models, including Linear Regression, Random Forest, Gradient Boosting Machines, Support Vector Regressors, Decision Tree Regressors, K-Nearest Neighbors, and LSTM networks, on a standardized time series dataset. Our methodology encompasses the preprocessing of time series data to address issues of non-stationarity and missing values, followed by the normalization of data to ensure model comparability. A key component of our approach is the use of cross-validation, specifically the K-Fold method, to rigorously evaluate the generalization capability of each model.

Our research makes several contributions to the field of time series forecasting. Firstly, it provides empirical evidence regarding the performance of a wide array of machine learning and deep learning models on time series data, thereby offering valuable insights into their strengths and limitations. Secondly, by employing a consistent evaluation framework, our study facilitates a direct comparison of these models, highlighting their relative efficacy in capturing temporal dynamics. Lastly, our findings contribute to narrowing the existing knowledge gap regarding the application of advanced neural network-based models, such as LSTMs, in the realm of time series forecasting, an area that has witnessed burgeoning interest but also presents significant challenges.

The remainder of this article is structured as follows: The next section delves into the methodology employed in our study, detailing the data preprocessing steps, model implementation, and evaluation criteria. This is followed by a presentation of our experimental results, where we discuss the forecasting performance of each model and provide a comparative analysis. Subsequent sections explore the implications of our findings for both theory and practice, address the limitations of our study, and outline avenues for future research. Finally, we conclude by summarizing the key contributions of our research and its relevance in the broader context of time series analysis and forecasting. Through this comprehensive exploration, our study endeavors to shed light on the effectiveness of various forecasting methodologies in navigating the complexities of time series data, thereby offering a roadmap for future research and application in this dynamic field.

2. RESEARH METHODS

2.1. Data Preprocessing

Our methodology begins with the preprocessing of time series of climate change data, which is crucial for preparing the dataset for effective model training and evaluation, the dataset can be collected from [27]. The dataset under consideration originates from a diverse range of fields, containing time series data that starts from the 9th column onwards. The initial step involves handling missing values, a common issue in time series datasets. We employ a technique to replace NaN values with zeros, acknowledging that alternative imputation methods may be more suitable depending on the specific characteristics of the data and the nature of missingness.

Following the treatment of missing values, the data undergoes a transformation to ensure it conforms to a floating-point format, facilitating numerical computation. The next crucial step in the preprocessing phase is the normalization of the dataset. Given the potential for wide variance in the magnitude of data points across the time series, normalization is applied to scale the data within a specified range. We utilize the MinMaxScaler, adjusting the data values to fall within the range [0, 1]. This normalization process is essential for models that are sensitive to the scale of input features, ensuring that no particular feature unduly influences the model's predictions due to its scale.

2.2. Feature Engineering and Dataset Preparation

Feature engineering is a pivotal aspect of our methodology, involving the transformation of the time series data into a structured format suitable for machine learning models. We create a lagged feature dataset, where each instance (X) represents the data at time t-1, and the target variable (y) corresponds to the data at time t, focusing on predicting the first column of the subsequent time step. This approach transforms the time series forecasting problem into a supervised learning task, enabling the application of various machine learning models. The dataset is then divided into two distinct formats: one suited for traditional machine learning models, reshaped into a 2D array (samples, features), and another prepared specifically for LSTM models, maintaining a 3D array structure (samples, timesteps, features) to accommodate the LSTM's input requirements.

2.3. Model Selection and Implementation

Our study encompasses a broad spectrum of forecasting models, ranging from traditional machine learning algorithms to advanced neural networks. The models include Linear Regression, Random Forest, Gradient Boosting Machines (GBMs), Support Vector Regressors (SVR), Decision Tree Regressors, K-Nearest Neighbors (KNN), and Long Short-Term Memory (LSTM) networks. Each model is selected for its unique characteristics and potential applicability to time series forecasting. The implementation of these models involves the use of standard libraries such as scikit-learn for machine learning algorithms and TensorFlow/Keras for LSTM networks. The models are configured with default parameters or minimal customization, acknowledging that hyperparameter tuning could further optimize their performance.

2.4. Training Strategy and Cross-Validation

A key component of our methodology is the training strategy, which employs K-Fold cross-validation to assess the generalizability of each model. Specifically, we utilize a 5-fold cross-validation approach, where the dataset is partitioned into five subsets, and the model training and evaluation are performed five times, each time using a different subset as the test set and the remaining subsets for training. This method ensures a robust evaluation by mitigating the impact of data variability on model performance.

2.5. Evaluation Metrics

The primary metric for evaluating model performance is the Root Mean Squared Error (RMSE), a widely used measure that quantifies the difference between the predicted values and the actual values. The RMSE is particularly informative in the context of forecasting, as it provides a direct interpretation of the prediction accuracy in the same units as the target variable. The use of RMSE facilitates a straightforward comparison of model efficacy across the different algorithms implemented in our study.

3. RESULT AND DISCUSSION

The comparative analysis of various machine learning and deep learning models for time series forecasting, as presented in the table 1, highlights intriguing insights into the performance of different methodologies as quantified by the Cross-Validation Root Mean Squared Error (CV RMSE). This analysis encompasses a broad spectrum of models, including the Long Short-Term Memory (LSTM) network, Linear Regression, Random Forest Regressor, Gradient Boosting Machine (GBM), Support Vector Regressor (SVR), Decision Tree Regressor, and K-Nearest Neighbors (KNN) Regressor.

Tabel 1. Models Comparison

Methods	Cross Validation RMSE
LSTM	**0.15523484782165892**
Linear Regression	5647815244.906905
Random Forest Regressor	0.15893855242272897
Gradient Boosting Machine (GBM)	0.16407510395166042
Support Vector Regressor (SVR)	0.15921804780414603
Decision Tree Regressor	0.1986917456091024
K-Nearest Neighbors (KNN)	0.16553573243649133

The LSTM model showcased superior performance with the lowest average CV RMSE of 0.155, underscoring the efficacy of recurrent neural networks in capturing the temporal dependencies and dynamics inherent in time series data. This outcome reinforces the notion that LSTM's architecture, particularly its ability to remember information for long periods, is highly suitable for forecasting tasks where historical data plays a crucial role in predicting future values. Linear Regression, however, exhibited an exceptionally high average CV RMSE of approximately 5.65 billion, indicating a substantial deviation between the predicted and actual values. This result suggests that Linear Regression, in its standard form, may not be well-suited for this specific time series dataset, potentially due to its inability to model complex, non-linear relationships inherent in the data.

The ensemble models, Random Forest and GBM, demonstrated commendable forecasting accuracy with average CV RMSE scores of 0.159 and 0.164, respectively. These results highlight the robustness of ensemble methods in time series forecasting, attributed to their ability to reduce overfitting and capture nonlinear patterns through the aggregation of multiple decision trees. Furthermore, SVR also showed promising results with an average CV RMSE of 0.159, closely competing with the ensemble models. This performance attests to the SVR's capability in handling non-linear relationships through its kernel functions, making it a viable option for time series forecasting. Next, The Decision Tree Regressor recorded an average CV RMSE of 0.199, the highest among the models excluding Linear Regression. This outcome might reflect the model's susceptibility to overfitting, where it captures noise in the training data that does not generalize well to unseen data. Lastly, the KNN Regressor, with an average CV RMSE of 0.166, showed that instance-based learning could also provide reasonable accuracy in forecasting time series data, emphasizing the importance of leveraging the similarities between temporal data points.

The evaluation presents a nuanced perspective on the applicability and performance of various models in time series forecasting. The stark contrast between the performance of LSTM and traditional Linear Regression highlights the critical need for models that can capture and leverage the temporal structure of time series data. While Linear Regression's poor performance might be attributed to its simplicity and linear assumptions, the success of LSTM and ensemble models underscores the significance of addressing the complex, non-linear interdependencies within time series data. The competitive performance of ensemble models and SVR further illustrates the potential of these approaches in managing the intricacies of time series forecasting. These models' ability to mitigate overfitting while capturing subtle patterns in the data makes them strong candidates for forecasting applications across various domains.

The relatively higher RMSE observed for the Decision Tree Regressor reinforces the notion that while capable of capturing complex relationships, decision trees must be carefully regularized or combined into ensemble methods to prevent overfitting and ensure generalizability. In light of these findings, it is evident that no single model universally outperforms others across all time series forecasting tasks. The choice of model should be guided by the specific characteristics of the dataset, including the nature of the temporal dependencies, the presence of non-linearities, and the overall complexity of the data. Future research may explore hybrid models that combine the strengths of different approaches, such as integrating LSTM with ensemble methods, to further enhance forecasting accuracy.

4. CONCLUSION

The study's findings contribute valuable insights into the selection and application of machine learning models for time series forecasting. The superior performance of the LSTM network highlights the potential of deep learning techniques in this domain, while the competitive results obtained from Random Forest and SVR models underscore the effectiveness of ensemble and kernel-based methods. The poor performance of the Linear Regression model serves as a reminder of the limitations of simpler models in capturing the complexity of time series data. In conclusion, this research underlines the importance of model selection in time series forecasting, emphasizing the need to consider the specific characteristics of the data and the underlying temporal dynamics. Future research could explore the integration of hybrid models, combining the strengths of different approaches to further enhance forecasting accuracy. Additionally, the exploration of advanced data preprocessing techniques, hyperparameter tuning, and the incorporation of external variables could offer new avenues for improving model performance. Through continued investigation and innovation, the field of time series forecasting stands to make significant advancements, offering increasingly sophisticated tools for data analysis and decision-making in various domains.

DAFTAR PUSTAKA

[1] F. Zennaro *et al.*, "Exploring machine learning potential for climate change risk assessment," *Earth-Science Rev.*, vol. 220, p. 103752, 2021.

[2] 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.

[3] P. D. Dueben, M. G. Schultz, M. Chantry, D. J. Gagne, D. M. Hall, and A. McGovern, "Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook," *Artif. Intell. Earth Syst.*, vol. 1, no. 3, p. e210002, 2022.

[4] Z. A. Mani and K. Goniewicz, "Adapting disaster preparedness strategies to changing climate patterns in Saudi Arabia: A rapid review," *Sustainability*, vol. 15, no. 19, p. 14279, 2023.

[5] O. H. Orieno, N. L. Ndubuisi, V. I. Ilojianya, P. W. Biu, and B. Odonkor, "The future of autonomous vehicles in the US urban landscape: a review: analyzing implications for traffic, urban planning, and the environment," *Eng. Sci. \& Technol. J.*, vol. 5, no. 1, pp. 43–64, 2024.

[6] R. S. Bang *et al.*, "An integrated chemical engineering approach to understanding microplastics," *AIChE J.*, vol. 69, no. 4, p. e18020, 2023.

[7] Y. Zhao, P. Deng, J. Liu, X. Jia, and J. Zhang, "Generative Causal Interpretation Model for Spatio-Temporal Representation Learning," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 3537–3548.

[8] H. Han *et al.*, "Advanced series decomposition with a gated recurrent unit and graph convolutional neural network for non-stationary data patterns," *J. Cloud Comput.*, vol. 13, no. 1, p. 20, 2024.

[9] V. Kumar, N. Kedam, K. V. Sharma, D. J. Mehta, and T. Caloiero, "Advanced machine learning techniques to improve hydrological prediction: A comparative analysis of streamflow prediction models," *Water*, vol. 15, no. 14, p. 2572, 2023.

[10] W. Leal Filho *et al.*, "Deploying artificial intelligence for climate change adaptation," *Technol. Forecast. Soc. Change*, vol. 180, p. 121662, 2022.

[11] M. E. Mondejar *et al.*, "Digitalization to achieve sustainable development goals: Steps towards a Smart Green Planet," *Sci. Total Environ.*, vol. 794, p. 148539, 2021.

[12] N. Rane, "Contribution of ChatGPT and other generative artificial intelligence (AI) in renewable and sustainable energy," *Available SSRN 4597674*, 2023.

[13] M. El Hajj and J. Hammoud, "Unveiling the influence of artificial intelligence and machine learning on financial markets: A comprehensive analysis of AI applications in trading, risk management, and financial operations," *J. Risk Financ. Manag.*, vol. 16, no. 10, p. 434, 2023.

[14] N. Schneider, "Unveiling the anthropogenic dynamics of environmental change with the stochastic IRPAT model: A review of baselines and extensions," *Environ. Impact Assess. Rev.*, vol. 96, p. 106854, 2022.

[15] F. K. Karim, D. S. Khafaga, M. M. Eid, S. K. Towfek, and H. K. Alkahtani, "A Novel Bio-Inspired Optimization Algorithm Design for Wind Power Engineering Applications Time-Series Forecasting," *Biomimetics*, vol. 8, no. 3, p. 321, 2023.

[16] K. E. ArunKumar, D. V Kalaga, C. M. S. Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, "Forecasting the dynamics of cumulative COVID-19 cases (confirmed,

recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA),” *Appl. Soft Comput.*, vol. 103, p. 107161, 2021.

[17] D. Xiao, J. Su, and others, “Research on stock price time series prediction based on deep learning and autoregressive integrated moving average,” *Sci. Program.*, vol. 2022, 2022.

[18] V. I. Kontopoulou, A. D. Panagopoulos, I. Kakkos, and G. K. Matsopoulos, “A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks,” *Futur. Internet*, vol. 15, no. 8, p. 255, 2023.

[19] A. Pasini and S. Amendola, “A Neural Modelling Tool for Non-Linear Influence Analyses and Perspectives of Applications in Medical Research,” *Appl. Sci.*, vol. 14, no. 5, p. 2148, 2024.

[20] E. S. Solano and C. M. Affonso, “Solar Irradiation Forecasting Using Ensemble Voting Based on Machine Learning Algorithms,” *Sustainability*, vol. 15, no. 10, p. 7943, 2023.

[21] H. Wan, S. Guo, K. Yin, X. Liang, and Y. Lin, “CTS-LSTM: LSTM-based neural networks for correlatedtime series prediction,” *Knowledge-Based Syst.*, vol. 191, p. 105239, 2020.

[22] E. Orsetti, N. Tollin, M. Lehmann, V. A. Valderrama, and J. Morató, “Building resilient cities: climate change and health interlinkages in the planning of public spaces,” *Int. J. Environ. Res. Public Health*, vol. 19, no. 3, p. 1355, 2022.

[23] A. M.-A. Qadir and A. O. Fatah, “Platformization and the metaverse: Opportunities and challenges for urban sustainability and economic development,” *EAI Endorsed Trans. Energy Web*, vol. 10, no. 1, 2023.

[24] S. Agrahari and A. K. Singh, “Concept drift detection in data stream mining: A literature review,” *J. King Saud Univ. Inf. Sci.*, vol. 34, no. 10, pp. 9523–9540, 2022.

[25] F. Bayram, B. S. Ahmed, and A. Kassler, “From concept drift to model degradation: An overview on performance-aware drift detectors,” *Knowledge-Based Syst.*, vol. 245, p. 108632, 2022.

[26] P. B. Weerakody, K. W. Wong, G. Wang, and W. Ela, “A review of irregular time series data handling with gated recurrent neural networks,” *Neurocomputing*, vol. 441, pp. 161–178, 2021.

[27] tarunrm09, “Climate Change Indicators.” 2021.