**Journal of Information Systems Engineering and Management**
2025, 10(24s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Advancing Information Systems for Smart Decision Making Using Machine Learning-Based Weather Prediction

Daniyal Kair[1], Ayan Baktygaliyev[2], Anara Kassymova[3], Alibek Joldybayev[4], Nazkey Ramazan[5], Amandyk Kartbayev[6]
[123456] *Kazakh-British Technical University, Almaty, Kazakhstan*

**ARTICLE INFO**

Received: 21 Dec 2024

Revised: 15 Feb 2025

Accepted: 28 Feb 2025

**ABSTRACT**

Accurate weather forecasting plays a crucial role in decision-making processes across various sectors. Traditional meteorological models often struggle with challenges such as missing data, insufficient spatial resolution, and inadequate assimilation of observational data. This study explores the integration of machine learning into weather prediction to enhance information systems for smart decision-making. The primary objective is to develop a robust machine learning-based approach for weather temperature prediction using open weather data. This research addresses missing data challenges, optimizes predictive accuracy, and establishes new methodologies applicable to Kazakhstan's meteorological systems. The study utilizes the rp5 weather database, which provides global weather data collected every three hours. Several machine learning models, including XGBoost, CatBoost, Linear Regression, and Bayesian Ridge, were applied using the scikit-learn library. Model evaluation was conducted based on mean squared error (MSE) and feature importance analysis, including SHAP values. Among all tested models, CatBoost demonstrated superior predictive performance with an MSE of 0.240654. Further optimization using Grid Search Cross-Validation indicated that the default hyperparameters were optimal. Feature importance analysis identified key variables affecting temperature prediction, including dew point, humidity, and atmospheric pressure at sea level. The relevance of this research is underscored by the increasing need for accurate weather predictions in industries such as agriculture, energy, and disaster management. The integration of machine learning enhances traditional forecasting methods, making them more adaptable to local climate conditions. These findings contribute to the advancement of meteorological information systems, providing actionable insights that can be utilized for optimizing weather-dependent operations in Kazakhstan and beyond.

**Keywords:** machine learning, weather prediction, CatBoost, decision support systems, feature selection, open data.

## 1. INTRODUCTION

Weather forecasting is a critical component of modern information systems, influencing decision-making in agriculture, transportation, energy management, and disaster preparedness. Accurate and timely weather predictions help optimize operations, mitigate risks, and enhance public safety. Despite recent advancements in meteorological modeling, challenges persist in collecting, processing, and analyzing weather data efficiently. Traditional forecasting approaches rely on numerical weather prediction (NWP) models that often struggle with issues such as data sparsity, incomplete observations, and limited computational efficiency [1]. The integration of machine learning into weather forecasting has the potential to address these limitations, transforming meteorological data into actionable insights that improve smart decision-making.

A key challenge in weather forecasting is the effective utilization of open data from various meteorological stations. Open data sources offer extensive observational records, yet existing methods often fail to fully exploit their potential for predictive modeling [2]. This study focuses on developing a machine learning-based system that leverages open weather data to improve forecast accuracy and support data-driven decision-making. Unlike traditional models, machine learning techniques can identify complex patterns in large datasets, making them highly suitable for improving predictive accuracy [3].

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

The current weather forecasting approach in Almaty faces several limitations that hinder the accuracy of temperature and thunderstorm predictions. First, insufficient spatial resolution in meteorological models results in a failure to capture microclimate variations influenced by local topography [4]. Second, inadequate data assimilation techniques limit the incorporation of satellite imagery, radar observations, and ground measurements, leading to propagation errors in forecasts [5]. Third, traditional models often oversimplify atmospheric processes, particularly convective systems responsible for thunderstorm activity, reducing predictive reliability [6]. Lastly, the lack of localized modeling strategies neglects urban heat islands and terrain-specific influences that significantly impact weather dynamics in Almaty [7]. Addressing these gaps requires a novel forecasting approach that integrates high-resolution data sources, advanced computational techniques, and domain-specific feature engineering.

Previous research has demonstrated the effectiveness of open data in improving forecasting accuracy. Studies have shown that integrating open weather station data enhances predictions of temperature, humidity, and wind speed in urban areas [8]. Machine learning algorithms, such as autoregression and hybrid statistical models, have also been found to outperform traditional approaches by enabling multi-step forecasts with lower error rates [9]. Recent advancements in short-term wind speed prediction further highlight the potential of machine learning to refine meteorological analysis [10]. In thunderstorm forecasting, machine learning-based hybrid models have demonstrated improved accuracy over conventional techniques, underscoring the need for integrating data-driven approaches into operational meteorology [11].

Beyond traditional statistical techniques, deep learning and hybrid models have emerged as promising alternatives in weather forecasting. Deep learning-based weather prediction (DLWP) methods leverage artificial neural networks to analyze spatiotemporal dependencies within meteorological data [12]. Comparative studies indicate that deep learning models provide competitive forecasting performance relative to NWPs, particularly when processing large, heterogeneous datasets [13]. Moreover, hybrid forecasting systems that integrate numerical models with machine learning optimization techniques have shown improvements in both computational efficiency and predictive accuracy [14].

The growing role of Internet of Things (IoT) technologies in meteorology further reinforces the shift toward data-driven forecasting. Real-time weather monitoring systems incorporating IoT sensors have successfully demonstrated enhanced prediction capabilities by capturing localized atmospheric changes [15]. A study utilizing IoT-enabled logistic regression models illustrated the feasibility of near-real-time weather forecasting using distributed sensor networks [16]. This convergence of IoT and machine learning highlights the potential of smart weather information systems to improve predictive reliability and decision-making processes.

Researchers in China have explored DLWP in comparison with traditional NWPs, emphasizing neural network architectures, benchmark datasets, and predictive accuracy across different spatiotemporal scales [17]. Their findings suggest that deep learning provides valuable tools for analyzing time-series weather data, reinforcing the case for integrating AI-based techniques into meteorological information systems. Similarly, localized forecasting studies in Sri Lanka have underscored the importance of region-specific modeling strategies to address climate-driven variability [18]. These findings hold particular relevance for Almaty, given its unique geographic features and microclimatic influences.

Despite the demonstrated benefits of machine learning in meteorology, existing approaches often suffer from optimization challenges and limited adaptability for real-time forecasting [19]. Many current models exhibit overfitting tendencies when dealing with complex stochastic processes, reducing their generalizability. Additionally, the absence of standardized feature selection methodologies in meteorological machine learning research complicates model interpretability and deployment. Addressing these gaps requires an adaptive framework that integrates automated hyperparameter tuning, explainable AI techniques, and domain-specific feature selection strategies.

This study aims to develop a machine learning-based forecasting model tailored to the meteorological conditions of Almaty. The proposed system will leverage open weather data to train predictive models, optimize feature selection using statistical and AI-driven approaches, and assess model interpretability through SHAP analysis. The relevance of this research extends beyond theoretical advancements, offering practical applications in climate-sensitive industries such as agriculture, energy, and disaster management. By enhancing the accuracy of weather predictions, machine learning-based forecasting can support more effective resource allocation, improve infrastructure planning,

and mitigate the economic impact of extreme weather events. Moreover, the development of an intelligent meteorological information system can aid policymakers in implementing data-driven climate adaptation strategies.

This study contributes to the advancement of weather forecasting methodologies by integrating machine learning with open meteorological data. By addressing the limitations of traditional models and leveraging AI-driven analytics, the proposed approach enhances predictive accuracy, supports smart decision-making, and improves the overall reliability of weather information systems. The findings provide a foundation for further research into real-time applications, hybrid forecasting frameworks, and deep learning-based enhancements to meteorological modeling.

## 2. METHODS

Our research is based on the weather measurement database provided by LLC "Weather Schedule," which contains open-access global weather data updated every three hours. For this study, we extracted historical weather data for Almaty City to train and evaluate our predictive model. The initial dataset consisted of 29 columns, but only relevant features were selected for model training. Table 1 provides an overview of the variables used after preprocessing. Several columns were excluded due to their irrelevance to prediction, strong bias, or excessive missing values. Some features, such as cloud type, were removed due to their negligible impact on model performance, while others were excluded based on statistical redundancy.

Table 1 : The embedding's for each column of the dataset

| Features | Description |
|---|---|
| T | Air Temperature (degrees Celsius) at a height of 2 meters above the ground |
| Po | Atmospheric pressure at the station level (millimeters of mercury) |
| P | Atmospheric pressure reduced to mean sea level (millimeters of mercury) |
| Pa | Baric trend: change in atmospheric pressure over the last three ours (millimeters of mercury) |
| U | Relative humidity (%) at a height of 2 meters above the ground |
| DD | Wind direction (points) at an altitude of 10-12 meters above the Earth's surface, averaged over the 10-minute period immediately preceding the observation period |
| Ff | Wind speed at an altitude of 10-12 meters above the Earth's surface, averaged over the 10-minute period immediately preceding the observation period (meters per second) |
| ff10 | The maximum value of the wind gust at altitude 10-12 meters above the Earth's surface in the 10- minute period immediately preceding the observation period (meters per second) |
| ff3 | The maximum value of the wind gust at altitude 10-12 meters above the earth's surface in the period between deadlines (meters per second) |
| WW | Current weather reported from the weather station |
| Tn | Minimum air temperature (degrees Celsius) for the past period (no more than 12 hours) |
| Tx | Maximum air temperature (degrees Celsius) for the past period (no more than 12 hours) |
| VV | Horizontal visibility range (km) |
| Td | Dew point temperature at a height of 2 meters above the earth's surface (degrees Celsius) |

To determine the most significant features for weather prediction, we determined a correlation of the features using the Python library Pandas' built-in .corr() method. This method supports different correlation coefficients, including Spearman rank correlation, Pearson correlation, and Kendall Tau correlation. Figure 1 presents the flowchart, which guided our feature selection process. The retained features were: 'T' (temperature), 'Po' (sea level pressure), 'P' (station pressure), 'Pa' (absolute pressure), 'Ff' (wind speed), 'Tn' (minimum temperature), 'Tx' (maximum temperature), 'VV' (visibility), 'U' (humidity), and 'Td' (dew point). We prioritized features with pairwise correlations between -0.1 and 0.1, as well as those that improved overall model performance.

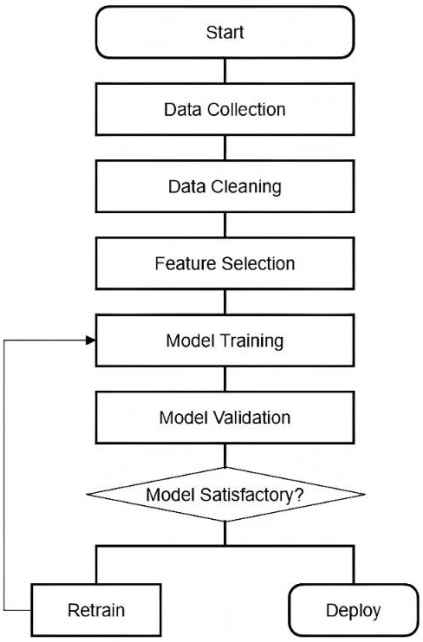


Figure 1: Overview of the guided feature selection process

After feature selection, we addressed missing values using data preprocessing techniques. Table 2 summarizes the completeness of the retained features. Only two columns contained fully populated data, while the others had missing values requiring treatment. For numerical attributes, replacing missing values with the mean yielded the best performance. Categorical attributes were converted into numerical equivalents to maintain consistency in the dataset. For instance, the dataset recorded "calm" as a textual value when wind speed was absent. Instead of discarding such rows, we replaced "calm" with zero, leading to improved predictive accuracy.

Table 2 : Number of rows with missing values for given attributes

| Column | Rows with NaN value |
|---|---|
| T | 0 |
| Po | 1 |
| P | 1 |
| Pa | 4 |
| Ff | 0 |
| Tn | 2568 |
| Tx | 2205 |
| VV | 1331 |
| U | 7 |
| Td | 7 |

Once the data was preprocessed, we split it into training and testing sets. The training dataset included weather records from April 1, 2022, to March 31, 2023, while the test dataset covered April 1, 2023, to March 31, 2024. This temporal division ensured that both sets represented a full annual cycle, maintaining seasonality consistency and enhancing the model's generalization ability. The chosen split method was essential for evaluating the model's predictive performance across different weather conditions, reducing potential overfitting to specific seasonal patterns.

The first step in our approach was to scale the dataset to enhance model performance. We utilized MinMaxScaler from the scikit-learn Python library, which normalizes feature values to a defined range—typically [0,1][0,1][0,1] or [−1,1][-1,1][−1,1] if negative values are present. Scaling improved model stability and overall prediction accuracy, as reflected in the final results.

The next step was selecting the best-performing general model for further experimentation. We evaluated multiple regression-based machine learning models during training and prediction phases, assessing their performance for weather forecasting. The following models were tested:

- **Gradient Boosting Regressor**: An ensemble learning method that combines multiple weak models (typically decision trees) by sequentially minimizing residual errors. It iteratively refines predictions, leading to strong overall performance.
- **Random Forest**: An ensemble method that constructs multiple decision trees on different subsets of the data. The final prediction is obtained by averaging outputs from all trees, reducing overfitting and improving robustness.
- **ElasticNet**: A hybrid of Lasso (L1) and Ridge (L2) regression, balancing feature selection and multicollinearity handling. It is particularly useful for high-dimensional datasets.
- **SGD Regressor**: A linear regression model optimized using Stochastic Gradient Descent (SGD), which updates parameters iteratively based on small data batches, making it suitable for large-scale datasets.
- **Bayesian Ridge**: A probabilistic linear regression method that applies Bayesian inference to determine regularization strength automatically, improving generalization.
- **Support Vector Regressor (SVR)**: A regression variant of Support Vector Machines (SVM) that finds an optimal hyperplane while allowing slight deviations, balancing prediction accuracy and flexibility.
- **CatBoost**: A gradient boosting algorithm designed for categorical features. It employs ordered boosting and specialized categorical encoding techniques to enhance performance.
- **Kernel Ridge Regression**: A combination of Ridge Regression and the kernel trick, enabling non-linear relationships to be captured by mapping data into a higher-dimensional space.
- **XGBoost**: An optimized gradient boosting algorithm known for its speed and efficiency. It builds sequential tree-based models while minimizing errors through advanced regularization techniques.
- **LightGBM**: A gradient boosting framework designed for efficiency, using a leaf-wise tree growth strategy that results in faster training and reduced memory consumption compared to conventional boosting methods.

Each model was assessed based on predictive accuracy, efficiency, and scalability. The results demonstrated that **CatBoost** outperformed the other models, making it the primary candidate for further optimization and refinement in our information system.

## 3. RESULTS

The results indicate that XGBoost and LightGBM exhibited slightly lower performance compared to CatBoost, suggesting that while gradient boosting algorithms are well-suited for this dataset, CatBoost outperformed its counterparts. Given that a single model was required for the final implementation, a hybrid approach was not pursued.

The next step involved tuning model parameters to improve performance. However, after extensive testing, it was found that hyperparameter tuning led to worse results compared to the default settings. Consequently, the default CatBoost model was retained without further optimization. Once the final model was trained and tested with the applied preprocessing techniques, it achieved an $R^2$ score of 0.97237, indicating high predictive accuracy. Since an $R^2$ score of 1.0 represents a perfect prediction, the obtained value suggests that the model performed exceptionally well. Table 2 presents a comparison of forecasting results before and after outlier removal, highlighting the impact of data preprocessing on model accuracy. The findings confirm that CatBoost performs significantly better without additional parameter tuning and is the most suitable model for weather prediction tasks [20].

Table 3: Model performance results after training the models

| Model name | Mean Squared Error |
|---|---|
| GB Regressor | 0.740616 |

| | |
|---|---|
| RF Regressor | 0.602757 |
| Linear Regression | 3.430981 |
| ElasticNet | 107.581766 |
| SGD Regressor | 2.586797 |
| Bayesian Ridge | 3.427670 |
| SVR | 2.530081 |
| CatBoost | 0.240654 |
| Kernel Ridge | 2.453900 |
| XGBoost | 0.678889 |
| LightGBM | 0.637373 |

With a mean squared error (MSE) of 0.240654, CatBoost outperformed all other tested models. To further validate its effectiveness, Grid Search Cross-Validation was applied to optimize hyperparameters. However, the results confirmed that parameter tuning did not enhance performance, reinforcing the decision to proceed with the default model (See Table 3).

Table 4: Sample of real and predicted temperatures comparison outputs

| **Time** | **Temperature in Actual Result** | **Temperature in Predicted Result** |
|---|---|---|
| 01.04.2023 23:00 | 2.0 | 1.899 |
| 01.04.2023 20:00 | 5.2 | 5.078 |
| 01.04.2023 17:00 | 8.4 | 8.459 |
| 01.04.2023 14:00 | 8.8 | 8.987 |

To better understand feature importance, we conducted tests using Feature Importance Analysis and SHAP (SHapley Additive exPlanations) values. These analyses revealed that the most significant features for the model were Dew Point (Td), Humidity (U), Atmospheric Pressure at Sea Level (P), and Atmospheric Pressure at Station Level (Po). SHAP analysis introduced additional influential features, including maximum (Tx) and minimum (Tn) temperatures [21]. The results indicate that temperature predictions for this region are primarily influenced by dew point, humidity, and atmospheric pressure at sea level, whereas atmospheric pressure at the station level and daily temperature extremes have a comparatively lower impact [22]. Other variables were found to have negligible influence on the predictive performance of the model.

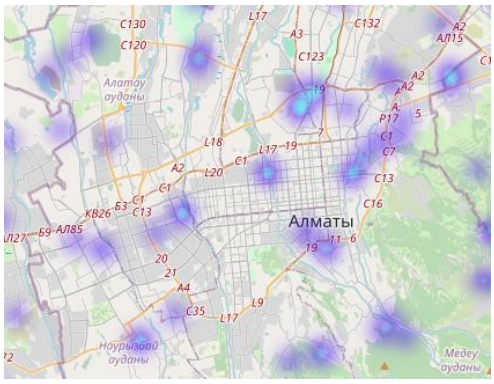


Figure 2 : Temperature anomaly outputs using Python scikitlearn library

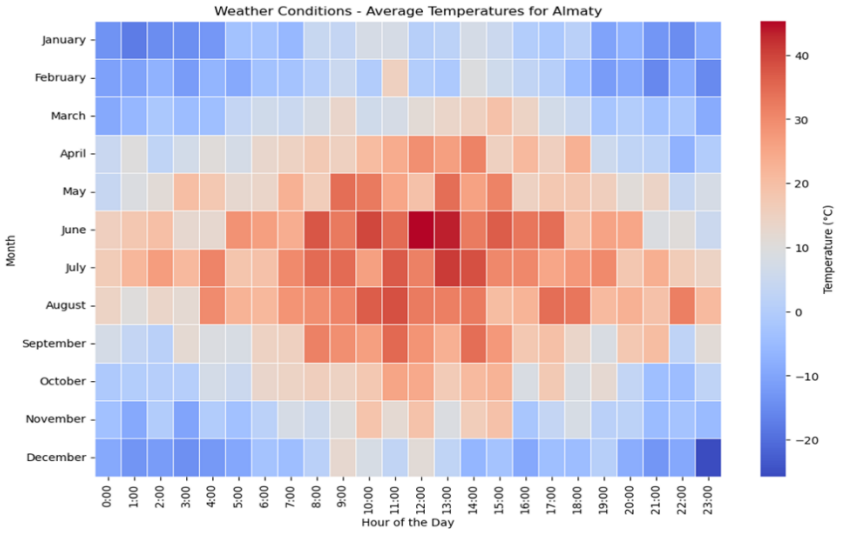


Figure 3 : Heatmap of predicted temperature outputs using Python scikitlearn library

For better visualization, Table 4, Figure 2, and Figure 3 provide a direct comparison between predicted and actual results over a short time period, showcasing the model's real-world forecasting accuracy. These findings reinforce the suitability of CatBoost as the optimal model for weather prediction, offering a highly accurate and computationally efficient approach without requiring complex hyperparameter tuning.

## 4. CONCLUSION

This paper provides a comprehensive framework for constructing a machine learning model tailored for numerical weather forecasting using publicly accessible datasets. The primary objective of this research was to explore and analyze the relationships between the features within these datasets and actual recorded weather data, focusing on how these relationships can enhance forecasting accuracy. Various machine learning techniques were evaluated to determine the best approaches for weather prediction in the Almaty city region, taking into account the unique regional characteristics, such as lower atmospheric pressure due to the mountainous topography, which influences local weather patterns.

Through this approach, we achieved a model that reliably forecasts weather based on the provided data. This model shows promising accuracy for Almaty and holds potential adaptability for use in predicting weather in other regions with similar or distinct climatological features. The research underscores the possibility of building scalable weather prediction applications capable of performing detailed numerical forecasting; however, technical constraints, particularly in processing power and computational resources, currently limit the speed and overall performance of the model. With improvements in computational infrastructure, especially increased CPU and GPU resources, we anticipate that model efficiency and forecast accuracy could be significantly enhanced, opening pathways for more responsive regional and future weather predictions in Almaty and beyond.

Several key areas for future research and improvements remain open. First, integrating deep learning techniques, such as Transformer-based models, could enhance the model's ability to capture long-term dependencies and complex temporal patterns in weather data. Incorporating real-time data streams from IoT-based weather monitoring systems would allow for near-instantaneous updates and improve forecast accuracy. These advancements will further improve the practical applicability of machine learning in meteorology, making predictions more precise, adaptable, and useful for decision-making in various sectors.

## REFERENCES

[1] Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature, 525*(7567), 47–55. https://doi.org/10.1038/nature14956

[2] Zhang, H., Liu, Y., Zhang, C., & Li, N. (2025). Machine Learning Methods for Weather Forecasting: A Survey. *Atmosphere*, *16*(1), 82. https://doi.org/10.3390/atmos16010082

[3] Islam, A., Attwood, S., Braun, M., Kamp, K., & Aggarwal, P. (2013). Assessment of capabilities, needs of communities, opportunities and limitations of weather forecasting for coastal regions of Bangladesh. *WorldFish*. http://dx.doi.org/10.13140/RG.2.1.1706.6485

[4] Pielke, R., & Uliasz, M. (1998). Use of meteorological models as input to regional and mesoscale air quality models—Limitations and strengths. *Atmospheric Environment*, 32(8), 1455–1466.

[5] Ravela, S., Emanuel, K., & McLaughlin, D. (2007). Data assimilation by field alignment. *Physica D: Nonlinear Phenomena, 230*(1–2), 127–145.

[6] Thompson, R. D. (1998). *Atmospheric processes and systems*. Psychology Press.

[7] Shao, W. (2016). Are actual weather and perceived weather the same? Understanding perceptions of local weather and their effects on risk perceptions of global warming. *Journal of Risk Research, 19*(6), 722–742.

[8] Hahn, C., Garcia-Marti, I., Sugier, J., Emsley, F., Beaulant, A. L., Oram, L., Strandberg, E., Lindgren, E., Sunter, M., & Ziska, F. (2022). Observations from personal weather stations.

[9] Chen, H., Zhang, Q., & Birkelund, Y. (2022). Machine learning forecasts of Scandinavian numerical weather prediction wind model residuals with control theory for wind energy. *Energy Reports, 8*, 661–668.

[10] Donadio, L., Fang, J., & Porté-Agel, F. (2021). Numerical weather prediction and artificial neural network coupling for wind energy forecast. *Energies, 14*(2), 338.

[11] Azad, M. A. K., Islam, A. R. M. T., Rahman, M. S., & Ayen, K. (2021). Development of novel hybrid machine learning models for monthly thunderstorm frequency prediction over Bangladesh. *Natural Hazards, 108*, 1109–1135.

[12] Wilgan, K., Rohm, W., & Bosy, J. (2015). Multi-observation meteorological and GNSS data comparison with numerical weather prediction model. *Atmospheric Research, 156*, 29–42.

[13] Kartbayev, A. (2015). Refining Kazakh word alignment using simulation modeling methods for statistical machine translation. In *Lecture Notes in Computer Science* (Vol. 9362, pp. 421–427). Springer. https://doi.org/10.1007/978-3-319-25207-0_38.

[14] Naveen, L., & Mohan, H. S. (2019). Atmospheric weather prediction using various machine learning techniques. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*.

[15] Kartbayev, A., Tukeyev, U., Sheryemetieva, S., Kalizhanova, A., & Uuly, B. K. (2018). Experimental study of neural network-based word alignment selection model trained with Fourier descriptors. *Journal of Theoretical and Applied Information Technology, 96*(13), 4103–4113.

[16] Krasnopolsky, V. M., & Fox-Rabinovitz, M. S. (2006). Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Networks, 19*, 122–134.

[17] Madan, S., Kumar, P., Rawat, S., & Choudhury, T. (2018). Analysis of weather prediction using machine learning & big data. *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. https://doi.org/10.1109/ICACCE.2018.8441683.

[18] Hennayake, K. M. S. A., Dinalankara, R., & Mudunkotuwa, D. Y. (2021). Machine learning-based weather prediction model for short-term weather prediction in Sri Lanka. 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS).

[19] Smatov, N., Kalashnikov, R., & Kartbayev, A. (2024). Development of context-based sentiment classification for intelligent stock market prediction. *Big Data and Cognitive Computing*, 8(51). doi: 10.3390/bdcc8060051

[20] Kair, D., & Kartbayev, A. (2024). A data-driven approach to assessing climate issues in coastal cities. BFT-2024, *BIO Web of Conferences, 130*, 06010. doi: 10.1051/bioconf/202413006010

[21] Verma, G., Mittal, P., & Farheen, S. (2020). Real-time weather prediction system using IoT and machine learning. 2020 6th International Conference on Signal Processing and Communication (ICSC). doi: 10.1109/ICSC48311.2020.9182766.

[22] Ren, X., Li, X., Ren, K., Song, J., Xu, Z., Deng, K., & Wang, X. (2021). Deep learning-based weather prediction: A survey. Big Data Research, 23. https://doi.org/10.1016/j.bdr.2020.100178