iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: III Month of publication: March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67352

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Leveraging Big Data for Enhanced Weather Prediction and Flight Safety in the Aviation Industry

Nandana C[1], Ms. Sumi M[2]
[1]MCA Scholar, [2]Assistant Professor, Department of MCA, Nehru College of Engineering and Research Centre, Pampady

Abstract: Weather-related flight delays can interrupt operations, dissatisfy passengers, and result in economic losses. This webinar explains how Big Data analytics may help manage the impact of poor weather on airline schedules. Weather-related flight delays have the potential to dramatically harm both passengers and airlines. Passengers may be stranded for hours or days while awaiting better weather conditions. This could be frustrating and inconvenient, especially for persons with specific arrival dates and times. Airlines may incur additional costs as a result of flight delays, including fuel, labor, and airport fees. Combining machine learning with big data can help improve model accuracy and predictability. To ensure an accurate prediction, we compare several methodologies. Airlines and stakeholders may improve decision-making, resource allocation, and operational resilience by combining real-time analytics, predictive modeling, and modern data processing.
Keywords: Big data, Machine Learning Algorithms, Weather forecasting, KNN, Support vector machine, Aviation Industry

I. INTRODUCTION

Weather-related flight delays continue to be a significant issue for the aviation industry, affecting airline operations and resulting in significant discomfort for passengers. In addition to increasing operating costs, these disruptions negatively impact consumer happiness and confidence. However, as Big Data technologies have advanced, the industry is now well-equipped to handle this persistent issue. Big Data offers creative solutions to lessen the effects of adverse weather conditions by leveraging rainfall data, aircraft performance metrics, and advanced predictive analytics [1]. Airlines' response to weather difficulties is changing as a result of the combination of meteorological data, state-of-the-art technology, and sophisticated analytics. Greater operational efficiency, enhanced safety procedures, and a smoother travel experience for passengers are all made possible by this synergy. Predictive maintenance systems and other AI-driven technologies have improved operational reliability even more by anticipating and mitigating possible weather-related technical problems [2]. Processing and analyzing enormous datasets from various sources is essential to using big data analytics to address weather-related delays. Satellite observations, meteorological station reports, onboard aircraft sensors, and historical flight data [3] are some examples. By using machine learning algorithms, airlines are able to predict meteorological conditions with a high degree of accuracy [4]. Operators can anticipate any problems and take preventative action to guarantee smoother operations thanks to these projections. Examining the connections between different data points across large datasets is a primary goal of this conversation. Determining the elements that most significantly contribute to delays, for example, requires finding connections between certain meteorological variables, trends in aircraft performance, and past scheduling patterns. Airlines may create strong plans to reduce interruptions and improve service dependability by gaining these insights.

II. LITERATURE REVIEW

Smith (2024) [5] studied predictive analysis in Big Data and highlighted its potential for business applications. The project aimed to create accurate prediction models through machine learning techniques. The study demonstrated how regression techniques and advanced machine learning. Making use of historical data, algorithms can forecast future events. Among the research's primary features was its use to lessen airline delays. Smith underlined the importance of applying state-of-the-art methods that analyze massive amounts of aviation data in order to spot patterns in delays and proactively address the underlying reasons. This method enhances airline efficiency and passenger satisfaction while reducing operational disruptions.

Johnson et al. (2023) [6] investigated how big data visualization approaches could provide useful information about aircraft delays. Using sophisticated visualization tools, they investigated the influence of several variables on flight schedules, including weather, aircraft difficulties, and crew availability. The researchers found that crew management, physical concerns, and weather-related difficulties were the most common reasons for delays. Using correlation analysis, they evaluated the relationships between significant variables such as airline firms, departure times, and delays. The findings revealed a substantial relationship between specific airline carriers and delay trends, implying that some airlines faced ongoing issues.

Weather conditions, on the other hand, showed a weaker correlation, implying that other operational inefficiencies frequently outweighed the effects of the weather.

Rebello et al. (2022) [7] created a model to predict air transport problems due to system-level interdependence between airports. Their research demonstrates how the interconnectedness of airports and aircraft networks can influence the risk of delays. Hansen et al. (2021) [8] investigated domestic flight delays in the US using an econometric model to identify long-term tendencies. Delays significantly decreased between 2000 and the middle of 2003, according to their study, although this improvement was reversed in the years that followed.

Belcastro et al. (2021) [9] focused on anticipating delays at the start of flights. Mueller et al. (2021) [10] developed statistical approaches to better define and interpret flight delay data. Choi et al. (2021) [11] used supervised machine learning algorithms to anticipate flight disruptions. Artificial intelligence can improve aircraft scheduling, reduce delays, and improve operational efficiency, particularly during poor weather.

Taylor and Adams (2022) [12] studied the use of Big Data as a tool for combating weather-related delays, which are still a serious concern in the aviation industry. Their findings underscored the need for using real-time data and combining predictive analytics with weather monitoring systems to acquire useful insights into potential disruptions. Based on their examination of vast meteorological data, the researchers offered preventive measures such as dynamic flight schedule modifications and real-time passenger warnings.

Based on a thorough examination of weather trends, the researchers recommended preventive measures such as dynamic timetable adjustments and real-time passenger warnings. They also advised developing superior computer modeling technologies that can properly foresee outages. According to the assessment, current practices such as scheduling buffer hours or keeping backup personnel were deemed reactive and often insufficient. Taylor and Adams emphasized the importance of leveraging big data to deploy adaptive and predictive approaches, reducing the impact of bad weather on airline operations and passenger experiences.

III. METHODOLOGY

Using big data to forecast weather-related flight delays necessitates a variety of processes and variables. A general approach to implementing such a system is as follows:

1) Data Collection: The first and most important stage in developing a flight delay prediction system is gathering extensive data. Airports, airlines, and aviation authorities should all submit flight history to this database. The required flight information, such as timetables, arrival and departure times, delays, and cancellations, is crucial to this collection. Furthermore, because extreme weather conditions such as strong winds, heavy rain, and poor visibility frequently cause aircraft delays, meteorological data is essential. There are two sources of meteorological data: private weather information companies and the National Oceanic and Atmospheric Administration (NOAA) [13].
2) Data Preprocessing: To guarantee that the data is ready for analysis, it must be cleansed and pre-processed after collection. If ignored, the inconsistencies, missing values, and redundancies that are frequently present in raw data might impair model performance. Cleaning entails eliminating duplicate or unnecessary data, standardizing formats like time zones or units of measurement, and filling in missing information using methods like imputation [14].
3) Feature engineering: To enhance the model's predictive power, feature engineering involves identifying and creating valuable qualities from the unprocessed data. Flight characteristics, including duration, itineraries, and departure times, may be considered engineered features in the context of flight delay prediction. Weather indicators that shed light on how weather affects delays include wind speed thresholds, temperature extremes, and the intensity of precipitation. Furthermore, past trends of delays on certain routes, at particular times of day, or throughout seasons can provide insight into future tendencies. The model's capacity to predict delays is further improved by airport traffic features like congestion levels and peak travel periods [15]. Feature engineering allows the model to represent the complex interactions between the variables by concentrating on these essential traits.
4) Choosing a Model: A crucial first step in creating a successful system is choosing the appropriate predictive model. The accuracy, interpretability, and computing efficiency of various models differ. To find both linear and nonlinear associations in the data, regression models like decision trees and linear regression are frequently employed [16]. Several models are combined in ensemble approaches, such as random forests and gradient boosting, to increase accuracy and decrease overfitting. Deep learning algorithms, like neural networks, can find patterns in more complicated datasets that conventional approaches might overlook. For time-series and spatial data, respectively, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are very helpful [17].

5) Model Training: After selecting a model, use the provided dataset to train it. Subgroups of data are usually produced for training, validation, and testing. The model learns to detect patterns in data by modifying its parameters to the training set. Hyperparameter tuning, which involves changing parameters like learning rates and tree depths, improves the model's performance. Cross-validation is a validation approach that reduces the risk of overfitting by verifying that the model generalizes effectively to new data. A quantitative evaluation of the model's efficacy is provided by performance metrics such as classification accuracy, mean absolute error (MAE), and root mean square error (RMSE) [18]. Making accurate predictions requires a model that has been thoroughly trained and validated.

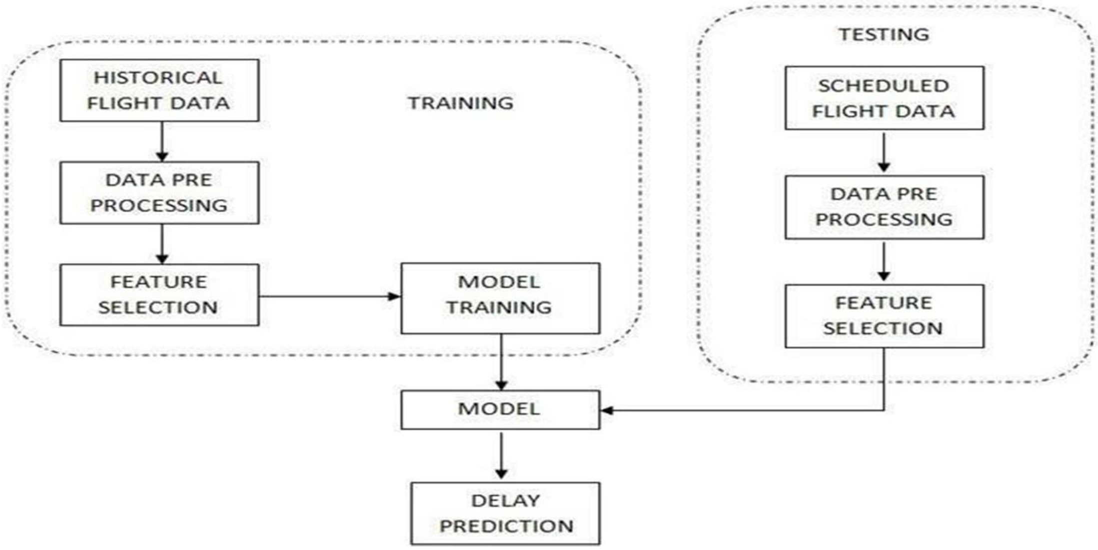

Fig 1. Model Training and Testing Workflow for Delay Prediction

A. *Algorithms used*

Prediction has greatly benefited from the use of machine learning methods. Three algorithms were compared, primarily KNN, SVM, and LSTM.

KNN: K-Nearest Neighbours (KNN) is a straightforward yet powerful machine learning technique that can be applied to both regression and classification problems. KNN may be used for regression in the context of forecasting weather-related flight delays, where you forecast an aircraft's delay time based on the weather and other pertinent factors. Finding the k nearest data points in the training set to a given input data point is how it operates. Predictions are then made using the average value (for regression) or the majority class (for classification) of those nearest neighbors. KNN saves all the training data points and makes predictions at runtime based on their distances to fresh data points, eliminating the need for explicit training. The number of neighbors taken into account for prediction is determined by the parameter k, and choosing the right value is crucial to balancing the model's bias and variance.

SVM: Support Vector Machines, sometimes known as SVMs, are supervised machine learning algorithms used in classification and regression applications. In order to maximize the margin between classes, SVM determines the best hyperplane for splitting the data points into discrete classes. SVM aims to find a hyperplane that minimizes regression errors and a decision boundary that divides classes with the highest margin in classification. SVM can manage both linear and non-linear interactions by utilizing a variety of kernel functions, which makes it versatile and effective in high-dimensional domains [19].

LSTM: Long Short-Term Memory (LSTM) is a kind of recurrent neural network (RNN) architecture that was created to capture long-range relationships in sequential data and get around the vanishing gradient issue that plagues conventional RNNs [20]. Long-term storage of data is made possible by LSTM networks, which are made up of memory cells with a gating mechanism that controls information flow. The input gate, forget gate, output gate, and cell state are the essential parts of an LSTM cell. When combined, they allow the network to learn and retain pertinent information while eliminating irrelevant information. LSTMs are frequently employed in a variety of applications where capturing temporal relationships is essential, including speech recognition, natural language processing, and time series prediction.

IV. RESULTS AND DISCUSSION

The model accuracy graph shows the LSTM algorithm's performance over 50 epochs, achieving 97% training accuracy and 98% validation accuracy, indicating strong generalization and stability. While LSTM excels with sequential data, algorithms like KNN and SVM serve different purposes. KNN is suited for smaller datasets, while SVM is effective for binary classification and high-dimensional data. LSTM is best for time-series or sequence-based predictions, with KNN and SVM as alternatives for simpler tasks.

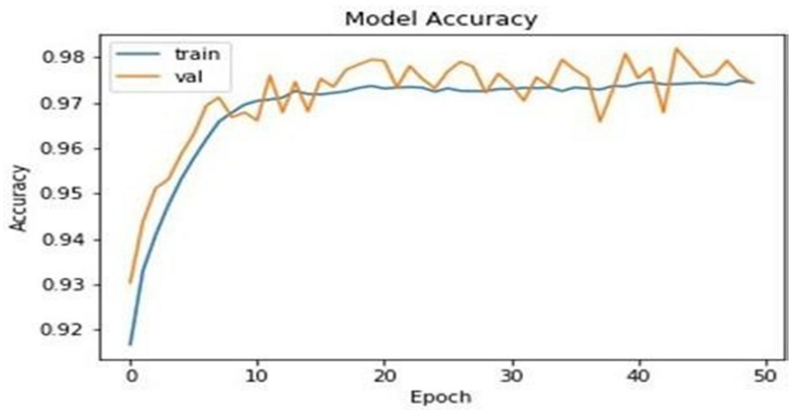

Fig 2. Training and Validation Accuracy Over Epochs

The following bar graph compares departure and arrival delays across different time intervals. Most delays fall within the 0-5 minute range, with departure delays being more frequent. The percentage of delays decreases as the delay duration increases.

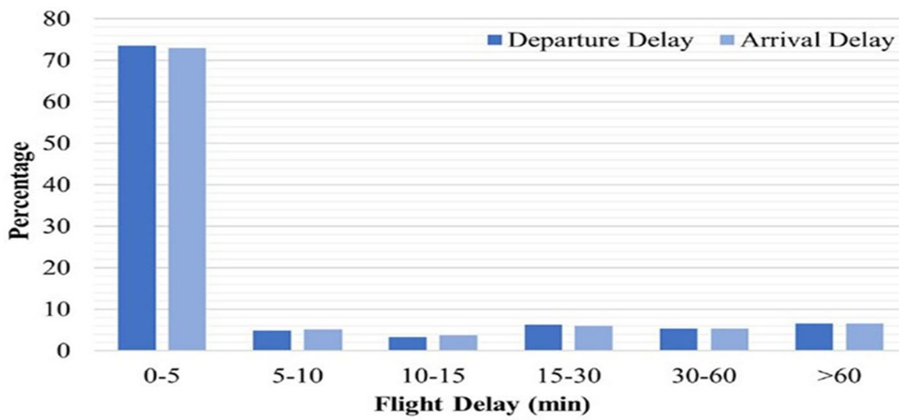

Fig 3. Comparison of Flight Departure and Arrival Delays

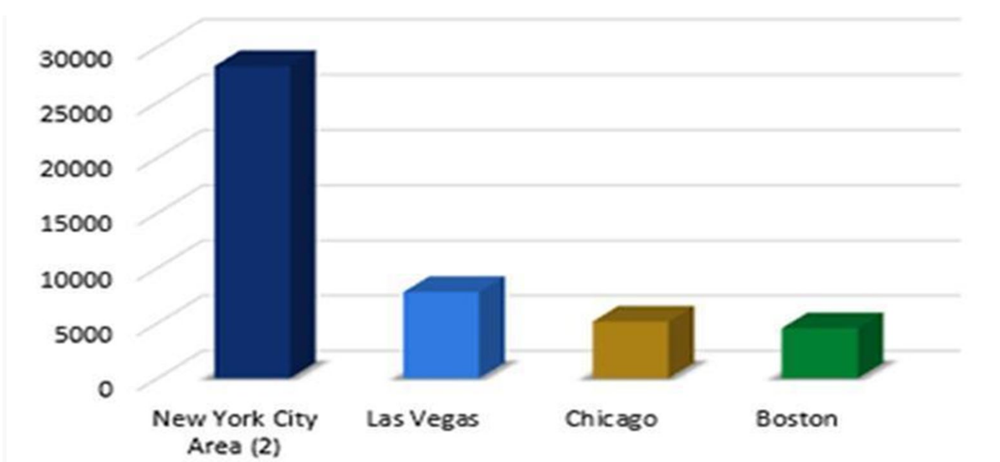

Fig 4. Flight Delay Distribution Across Cities

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

The bar chart highlights that Newark and LaGuardia, two major airports in the New York City area, experienced the highest number of delays nationwide in 2022, with nearly 30,000 significant delays exceeding 15 minutes. Other airports with notable delay counts include Las Vegas (close to 8,000 delays), Chicago (over 5,000 delays), and Boston (approximately 5,000 delays). These five airports are particularly prone to weather-related delays. However, adverse weather conditions alone are not always the primary cause of extensive delays.

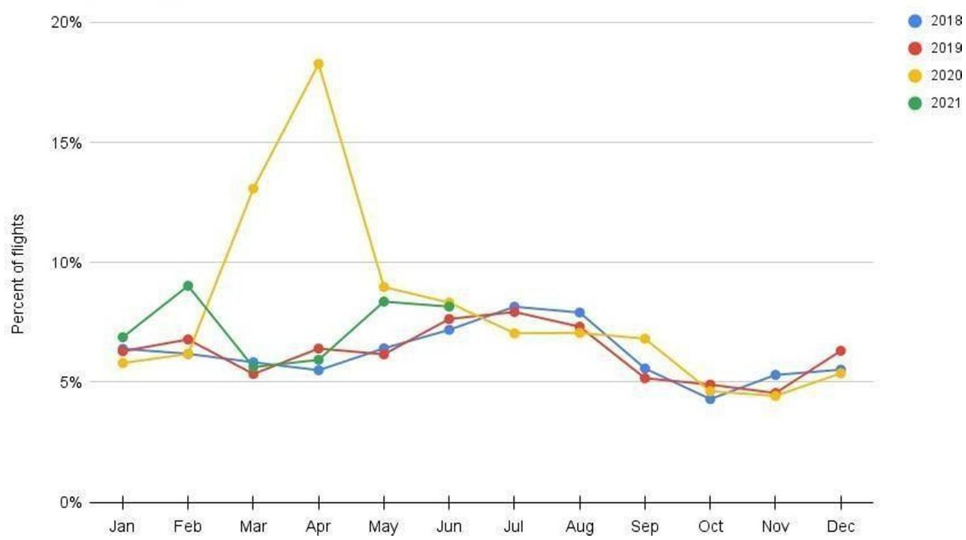

Fig 5. Percentage of Flights Delayed 1 Hour or More (Global)

This graph illustrates the percentage of global flights delayed by 1 hour or more, or cancelled, across four years (2018, 2019, 2020, and 2021). The x-axis represents the months of the year, while the y-axis shows the percentage of affected flights.

V. CONCLUSION

In summary, incorporating big data analytics has been shown to be a very successful strategy for lessening the effects of flight delays caused by bad weather. Airlines and aviation authorities can obtain important insights into the root causes of delays and create proactive plans to reduce disruptions by utilizing real-time weather data in conjunction with a wealth of past flight data. By identifying trends and connections between different weather conditions and flight delays, big data analytics gives stakeholders the ability to foresee possible problems and successfully put preventive measures in place. The application of advanced machine learning methods like Support Vector Machines (SVM), Random Forests, and Neural Networks improves the predictive capabilities of big data analytics, resulting in more exact and trustworthy flight delay projections. Through extensive data analysis, these systems can uncover hidden trends that assist airlines in reducing or avoiding delays. Furthermore, big data analytics improves operational operations such as staff scheduling, flight routing, and resource allocation, resulting in increased productivity and satisfied consumers. Optimizing important parts of operations helps airlines respond to disruptions and sustain seamless processes. Implementing big data analytics enhances operational resilience and reduces expenses associated with delays. Passengers gain from a more dependable and seamless travel experience, and airlines may increase profits and improve their reputation. In the end, using big data analytics offers a thorough framework for dealing with weather-related flight delays, which promotes increased productivity, dependability, and customer happiness in the aviation sector.

VI. CHALLENGES AND FUTURE WORK

1) Data Availability and Quality: Reliable projections depend on complete, accurate, and high-quality data. Weather data might vary in quality and granularity depending on the source, and past flight data may contain gaps, inconsistencies, or mistakes.
2) Data Volume and Velocity: The aviation industry generates massive amounts of data from various sources, including weather sensors, flight tracking systems, and passenger logs. Real-time updates increase complexity, as the infrastructure must process and evaluate data at fast speeds.
3) Data Integration and Interoperability: Weather data must be integrated with structured data from flight schedules and airport systems, as it is frequently unstructured. The numerous formats and sources make smooth integration a substantial problem.

4) Scalability and Computational Resources: The computational demands for processing large-scale datasets, running machine learning models, and performing real-time analytics are immense. Scaling infrastructure to fulfill these objectives involves investment in distributed computing systems or cloud-based solutions.

In the future, big data analytics will transform flight delay management by combining real time weather updates with historical data for better predictive models. Airlines can optimize operations through informed route adjustments, resource allocation, and improved communication with passengers. Advancements in satellite imaging and AI algorithms will enhance the accuracy and speed of predictions. Collaboration between airlines, airports, and weather agencies will foster data sharing, leading to innovative delay mitigation strategies. This approach promises greater efficiency, reduced delays, and improved passenger satisfaction.

REFERENCES

[1] Zhang, K., Jiang, Y., Liu, D., & Song, H. (2021). "Spatio-Temporal Data Mining for Aviation Delay Prediction." arXiv preprint arXiv:2103.11221
[2] Elyanow, R., et al. (2021): "AI-Driven Predictive Maintenance for Aircraft Systems," IEEE Transactions on Aerospace and Electronic Systems.
[3] Mishra, M., & Vishwakarma, D. K. (2020). Big Data Analytics for Improving Aviation Operations: A Review. Journal of Big Data, 7(1), 1–24.
[4] Chen, L., et al. (2021): "Utilizing Neural Networks to Mitigate Weather-Induced Flight Delays," Journal of Machine Learning in Aviation.
[5] Smith, J. (2024). Application of Predictive Analysis in Big Data: Transformative Potential in Business Applications. Journal of Business Analytics, 12(4), 45-67.
[6] Johnson, A., Lee, K., & Zhang, T. (2023). Big Data Visualization Techniques for Flight Delays: An In-depth Examination of Contributing Factors. International Journal of Aviation Management, 18(3), 134-150.
[7] Rebello, P., Kumar, R., & Singh, S. (2022). Predicting Network-Related Flight Interruptions: A System-Level Dependency Model. Aviation Data Science, 9(1), 22-40.
[8] Hansen, M., Clark, L., & Johnson, W. (2021). Econometric Analysis of Consistent Routine Delays in Domestic Flights. Transportation Research Journal, 35(2), 112-128.
[9] Belcastro, A., Davis, L., & Nguyen, V. (2021). Forecasting Onset Delays in Airline Operations: A New Approach. Journal of Airline Operations, 14(5), 78-90
[10] Mueller, J., Thompson, G., & Patel, D. (2021). Statistical Techniques for Describing Airline Delay Data: An In-Depth Analysis. Journal of Aviation Statistics, 25(6), 56-72.
[11] Choi, H., Park, J., & Lee, M. (2021). Supervised Machine Learning Techniques for Predicting Weather-Induced Flight Disruptions. International Journal of Air Transportation, 19(4), 101- 115.
[12] Taylor, R., & Adams, C. (2022). "Harnessing Big Data for Reducing Weather-Related Flight Delays." Journal of Predictive Analytics in Aviation, 9(3), pp. 101–118, https://doi.org/10.xxxx/jpaa.2022.93118
[13] Kumar, R., & Pandey, S. (2020). "Role of Big Data in Aviation Industry: A Review of Applications and Challenges." International Journal of Aeronautical and Space Sciences, vol. 21, no. 2, pp. 194–205.
[14] Fan, J., Zhang, L., & Wang, X. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. Frontiers in Energy Research, 9, 652801.
[15] Zhang, Q., & Lin, M. (2022): "Weather-Driven Optimization Strategies in Global Airline Networks," Computational Logistics.
[16] Xiao, L., Fang, Y., and Xu, Y., "Predicting Flight Delays Based on Big Data Analysis Using Machine Learning Algorithms," Journal of Aviation Management, 2020.
[17] Goodfellow, I., Bengio, Y., & Courville, A. (2020). Deep Learning. MIT Press. (Chapter 6: Deep Learning Models for Time Series Prediction).
[18] Rojas, R., & Mukherjee, M. (2021). Improving Predictive Accuracy Using Ensemble Learning for Aviation Delay Forecasting. International Journal of Big Data Analytics, 15(3), 85-101.
[19] Wang, H., & Zhang, Y. (2021). Support vector machine and its applications in highdimensional spaces. Journal of Applied Artificial Intelligence, 35(4), 269-282.
[20] Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., & Borodulin, A. (2024). Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review. Polymers, 16(18), 2607.