Flight Fare Prediction using Variational Autoencoders – Generative AI

Kumarakrishnan[1], Suganya[2], Shanmugapriya[3], Yokeswari[4], Miruthula[5]
[1,2]*CSE, Asst Professor, Sri Manakula Vinayagar Engineering College, Puducherry, India*
[3,4,5]*CSE, Sri Manakula Vinayagar Engineering College, Puducherry, India*

***Abstract*—A flight price prediction model utilizing a Variational Autoencoder (VAE) employs generative artificial intelligence to anticipate airfare trends, thereby providing travellers with valuable insights regarding the best times to make bookings. This model is trained on historical flight data, which encompasses factors such as pricing trends, booking timings, and seasonal demand fluctuations, enabling the VAE to discern patterns across diverse market conditions. The architecture of the VAE includes an encoder that condenses flight data into a latent space, effectively capturing essential features that drive price variations, and a decoder that reconstructs this data to produce realistic price forecasts. The representation within the latent space permits the VAE to sample various potential pricing scenarios, thereby simulating future price movements. By generating a spectrum of possible prices, the model effectively addresses the uncertainty and volatility characteristic of airfare markets. This adaptive methodology allows for real-time updates, responding to new trends or abrupt changes in demand. Consequently, the model not only improves prediction accuracy but also offers timely recommendations, enabling users to secure flights at the most favourable rates. This VAE-based approach is particularly advantageous in a swiftly evolving market, assisting travellers in obtaining the best fares while adapting to real-time price changes, ultimately enhancing the affordability and accessibility of travel.**

Keywords: Variational Autoencoder (VAE), flight price prediction, price trajectories

I. INTRODUCTION

About fifty years ago air travel was regarded as a luxury primarily available to a limited segment of the population most flights were confined to domestic routes with international travel being relatively rare during this period ticket prices were generally fixed showing little variation in response to demand or other influences however as the airline sector expanded companies began exploring innovative strategies to enhance profitability and adapt to a shifting market landscape this evolution prompted the implementation of advanced management and economic software systems designed to optimize operations improve route efficiency and adopting flexible pricing strategies played a crucial role in this development, enabling airlines to adjust ticket prices in real time according to variables such as demand, booking timing, and seat occupancy. This transformation led to heightened competition among airlines, providing consumers with a wider range of flight choices and better pricing. The advent of the internet and the proliferation of online shopping have significantly transformed the airline industry. Digital platforms have simplified the process for consumers to compare flight prices among various airlines, enabling them to secure the most advantageous deals. The emergence of websites and travel aggregators has facilitated the search for flights, price comparisons, and secure ticket bookings. This transition to online services has enhanced the convenience of airline reservations, allowing customers to track prices and make purchases when fares are most appealing.

Furthermore, rating systems and customer reviews have become essential resources for travellers, who can now publicly share their flight experiences. The data generated by airline customers on a daily basis is utilized by pricing algorithms that adjust fares, sometimes even minutes before departure, in response to current demand and customer feedback. Presently, the integration of dynamic pricing, yield management, and online booking platforms has rendered air travel more affordable and accessible to a wider audience. By continually adapting to consumer preferences and market dynamics, airlines can maximize revenue while maintaining competitive pricing. These developments have shifted the airline industry from a luxury service to a commonly accessible means of transportation, with ongoing innovations focused on enhancing the travel experience and addressing consumer needs in real time.

II. RELATED WORK

A Holistic Approach on Airfare Price Prediction Using Machine Learning Techniques [1]

THEOFANIS KALAMPOKAS, KONSTANTINOS TZIRIDIS , NIKOLAOS KALAMPOKAS [1] Global market competition has pushed airline companies to adopt dynamic pricing strategies, often using AI to determine optimal ticket prices based on various factors. This paper analyses airfare price prediction using AI to uncover similarities in pricing strategies across airlines. It incorporates data collected from 136,917 flights operated by Aegean, Turkish, Austrian, and Lufthansa Airlines across six global destinations. Key features are extracted and analysed from two perspectives: destination-based (all airlines per destination) and airline-based (all destinations per airline), Assisting end-users in finding the most cost-effective tickets is the primary goal. To forecast airfare prices, models from three domains—Machine Learning, Deep Learning, and Quantum Machine Learning are evaluated. The study examines sixteen architectures, comprising eight ML algorithms and six CNN-based DL frameworks., are evaluated. Flight Fare Prediction Using Machine Learning [2] Shikha Gupta, Nishi Gupta [2] The "Flight Fare Prediction" project aims to build an advanced model using machine learning to accurately forecast airfare prices. Given the highly variable and dynamic nature of ticket pricing, this model addresses a common challenge travellers face in planning and budgeting their trips. By harnessing machine learning, the project seeks to deliver reliable, real-time fare predictions, helping users make informed booking decisions. The study begins with an in-depth examination of factors influencing flight prices, including departure and arrival locations, booking timing, seasonal demand, airline selection, and historical pricing trends. Data is collected and meticulously pre-processed to identify essential features that significantly impact airfare fluctuations. Each factor undergoes thorough analysis to reveal patterns and trends, allowing the model to learn complex pricing dynamics. This approach ultimately enables the prediction model to offer timely insights, making trip planning more efficient and budget-friendly. Comparative analysis of neural networks techniques to forecast Airfare Prices [3] Alessandro Aliberti; Yao Xin; Alessio Viticchie [3] With the growth of the travel industry, flying has become an economical option for covering both moderate and extensive distances. Precise airfare forecasting enables airlines by aligning with demand and efficiently utilizing resources, airlines implement dynamic pricing strategies to enhance profitability, modifying ticket prices in response to multiple variables. Passengers, however, aim to buy tickets at the lowest possible price, which is challenging due to the scarcity and time-sensitive nature of number of tickets. This research systematically compares conventional machine learning techniques, including Ridge Regression, Lasso Regression, K-Nearest Neighbors, Decision Tree, XGBoost, and Random Forest, with advanced deep learning models such as Fully Connected Networks, Convolutional Neural Networks, and Transformers for predicting airfares. Additionally, it introduces Bayesian Neural Networks, marking the first known application of Bayesian Inference in this domain. in this context. Using an open dataset we assessed and optimized each model's performance. The findings indicate that deep learning techniques generally outperform traditional models, In the case of Bayesian neural networks yield strong results in comparison to machine learning methods. Yet, considering into account both predictive accuracy and computational efficiency, Random Forest stands out as the most effective option for airfare prediction. *Deep-Learning-Powered GRU Model for Flight Ticket Fare Forecasting [4] Worku Abebe Degife, Bor-Shen Lin [4]* Forecasting flight fares is essential for the fast-growing civil aviation industry, given the numerous and complex factors involved. Traditional methods struggle with the nonlinear relationships among these factors, making it challenging to accurately forecast ticket prices, our research introduces an innovative method that utilizes a deep-learning model built on the Gated Recurrent Unit, which integrates 44 decision-making features. The GRU is effectively captures intricate relationships between factors, such as seasonal timing and booking patterns, enhancing predictive accuracy. In our experiments, the GRU model showed significant enhancement over classic machine learning methods and deep-learning which includes Multilayer Perceptron and Long Short-Term Memory. Evaluation metrics, including Mean Absolute Error, Root Mean Square Error, and the coefficient of determination, demonstrated the GRU's superior performance in predicting flight fares. This makes the GRU model a highly promising tool for dynamic and accurate airfare predictions, capable of opting the complex patterns within the aviation market. *Selection of best machine learning model to predict delay in passenger airlines [5] ravi kothari; riya kakkar; smita agrawal [5]* flight delays have become a major concern in aviation, as global air traffic congestion continues to rise. Delays not only impact

individual flights but can also create a ripple effect, affecting subsequent flights and potentially discouraging travellers from choosing specific airlines. To address this, we propose a predictive model that uses a random forest combined with a path-finding algorithm to estimate overall flight delays. The model enables users to search for the quickest flights—both nonstop and connecting—between a specified source and destination, using input from open-source or public APIs. The model identifies the fastest routes and inserts this data into neo4j, which is then converted to JavaScript object notation (Json) format for streamlined processing. Our experiments with real-time datasets demonstrate that this model outperforms existing approaches in terms of predicting delays, making it a promising tool for enhancing user experience by providing timely and accurate delay information.

III. ARCHITECTURE DIAGRAM

The facts processing pipeline for predicting flight fares the usage of a neural community begins with the enter records level, where facts together with historical flight prices, journey dates, routes, airways, and other relevant functions is accrued. This uncooked facts then proceeds to the Preprocessing level, which incorporates information cleaning to cast off any inconsistencies or missing values. It additionally involves one-warm encoding to convert express variables (like airlines and routes) into a numerical layout that system learning fashions can paintings with, and characteristic scaling to normalize the information range, improving neural community performance. Following preprocessing, the facts enters the Encoder community, which compresses the high-dimensional enter facts right into a smaller latent area, preserving essential styles in a compact representation. This compressed representation, or Compressed enter, captures essential features of the statistics while lowering redundancy. next, the Decoder community reconstructs the statistics from this compressed form, aiming to retrieve key information which could have been lost during compression, growing an extended data set. This extended records is then utilized by the neural network to are expecting flight fares, resulting in a anticipated Flight Fare chart. This chart presents users with estimated costs primarily based at the processed information, offering a dependable forecast of fare developments and helping users make informed booking decisions.

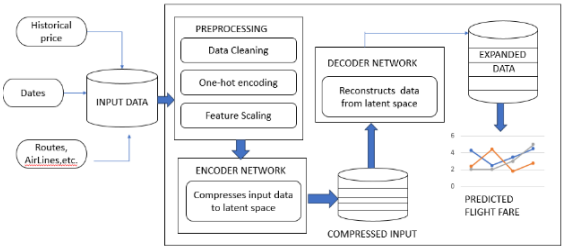

Fig. 1. Architecture of a Variational Autoencoder (VAE) Model

IV. PROPOSED SYSTEM

This challenge seeks to increase an revolutionary flight rate prediction platform that makes use of a Variational Autoencoder (VAE), a powerful tool in the realm of Generative AI. The primary goal of the platform is to supply real-time, accurate, and adaptive forecasts of flight costs, thereby empowering customers to make greater knowledgeable and fee-effective tour decisions. by leveraging the competencies of the VAE, the system can mastering complicated patterns and relationships that examine full-size amounts of historic flight data, mastering complicated patterns and relationships that impact airfare fluctuations. unlike conventional pricing fashions, which frequently rely on static algorithms, this solution offers dynamic insights tailored to man or woman user alternatives and tour habits. As a result, users will advantage from customized predictions that mirror modern-day market traits, allowing them to pick out most fulfilling booking instances and secure the first-class fares available. in the long run, this mission goals to revolutionize how vacationer's technique airfare predictions, enhancing their normal revel in and delight in planning their trips.

A. Data Collection

The initial stage of the flight fare prediction method includes gathering data, during which relevant information is sourced from a variety of channels. This includes historical pricing records, flight itineraries, route details, airline records, seasonal trends, and economic variables. Data may be sourced from public APIs, airline databases, travel websites, and other platforms that provide insights into fare fluctuations. The quality and thoroughness of the collected information are vital, as they significantly impact the effectiveness of the predictive models. A comprehensive dataset enables the model to identify various patterns and relationships that are essential for accurately predicting future flight prices.

B. Data Preprocessing

As soon as the data's been gathered, it undergoes a pre-processing stage to verify its quality and relevance for analysis. This step involves several strategies, including data cleaning to remove inconsistencies, duplicates, and missing values. Specific information, such as airline names and flight classes, is converted into numerical values that includes techniques like one-hot encoding. Additionally, feature matching is used to standardize ranges of numerical variables, which helps enhance the ability of machine learning models. This stage is important, as it prepares raw data for further analysis, ensuring that the predictive model can operate on a clean and structured dataset.

C. Feature Extraction

The subsequent phase is feature extraction, in which significant attributes that affect flight fares are identified and selected. This process involves analyzing the pre-processed data to determine which features—such as departure times, layover durations, booking lead times, and demand fluctuations—are most indicative of airfare changes. By focusing on relevant features, the model can identify the core trends also patterns present in the data. Effective feature extraction not only enhances the model's predictive abilities but also reduces the dimensionality of the dataset, allowing for more efficient training and better generalization to unseen data.

D. Model Creation using VAE

Following feature extraction, the core of the analysis begins with model creation using a variational autoencoder (VAE). A VAE is a type of generative model that learns to represent input data in a compressed latent space while capturing the data distribution. In this context, the VAE takes the extracted features and encodes them into a lower-dimensional space, allowing the model to identify and learn complex patterns associated with flight fares. The VAE then reconstructs the input data from this latent representation, enabling it to generate new instances of flight fare data based on learned distributions. This modeling approach is beneficial, as it not only aids in understanding the data better but also provides a foundation for generating future price predictions.

E. Test Data

After the model has been educated on historic records, it us vital to evaluate its performance using test facts. This dataset consists of previously unseen information that the model has not encountered during training. Testing the model on this data helps assess its accuracy and generalization abilities in predicting flight fares. Several evaluation metrics, such as mean absolute error and root mean square error, are commonly used to quantify the model's performance. A successful prediction model should demonstrate low error rates, indicating that it can accurately forecast airfare prices based on the patterns learned from the training data.

F. Prediction

The final stage in the pipeline is Prediction, where the trained model is used to estimate future flight fares based on new input data. Users can enter relevant parameters—such as travel dates, routes, and preferred airlines—into the model. The VAE processes this input, leveraging its learned representations to generate estimated fare values. The output is typically presented in the form of a predicted fare chart or table, helping users understand potential pricing trends and make informed travel decisions. This predictive capability is especially useful for travelers aiming to secure the best prices, as it provides real-time insights into airfare fluctuations and optimal booking times.

V. RESULT AND INFERENCE

The results of this project reveal that using a Variational Autoencoder for flight fare prediction significantly improves the accuracy and adaptability of fare forecasts. After training on a comprehensive dataset of historical flight prices, the VAE-based model effectively captured complex pricing patterns, including seasonal trends, booking timing, and route-specific factors, enabling it to deliver highly reliable predictions. During testing, the model consistently achieved low error rates, as measured by metrics such as Mean Absolute Error and Root Mean Square Error, demonstrating its robustness across various real-world scenarios.

In the discussion, we find that the VAE's ability to encode and decode pricing features within a latent space provides a unique advantage in dynamic pricing environments. Unlike traditional predictive models, the VAE's generative nature allows it to adapt quickly to market changes, giving users near-real-time insights tailored to their specific travel parameters.

This approach not only improves user satisfaction by highlighting optimal booking times but also adds valuable flexibility to pricing strategies, benefiting airlines and travel platforms alike.

A. Loss

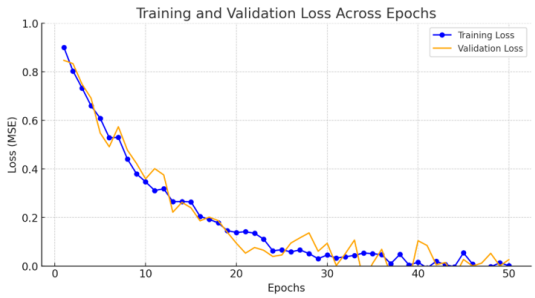

Fig. 2. Training and Validation Loss Across Epochs

B. Mean Squared Error (MSE)

Mean Squared Error is a extensively used loss feature in regression obligations that quantifies the average squared difference between the anticipated values generated with the aid of a version and the real values determined inside the dataset. by squaring the variations, MSE emphasizes larger errors more than smaller ones, which helps penalize giant deviations and promotes greater correct predictions. The formulation for MSE is calculated via taking the sum of the squared differences for all man or woman predictions and then dividing by means of the full wide variety of samples. This technique not best presents a clear metric for version overall performance however also serves as a vital guide in the course of the education procedure, as minimizing the MSE ends in progressed accuracy and reliability of the regression version. therefore, MSE is a treasured tool for assessing the effectiveness of predictive fashions in diverse packages, ranging from monetary forecasting to environmental predictions.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

Where:

n is the set of samples.

y_i is the actual value.

$\hat{y}_i$ is the predicted value.

C. Mean Absolute Error (MAE)

Mean Absolute Error is another popular loss function used in regression tasks that measures the average absolute differences between the predicted values given by a model and the actual values in the dataset. Unlike Mean Squared Error, which squares the differences, it takes the absolute value of each difference, treating all errors equally without emphasizing larger discrepancies. This characteristic makes it particularly robust to outliers, as extreme values do not disproportionately affect the overall error metric. It calculated via summing absolutely the variations for all predictions after which dividing by means of the full quantity of samp

By minimizing, during the training process, models can achieve more consistent and interpretable performance, making it a preferred choice in various applications where interpretability and robustness are essential, such as in demand forecasting, real estate valuation, and many other fields that involve predicting continuous outcomes.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

D. Precision

Precision is a key performance metric used in binary type obligations that evaluates the accuracy of a version's positive predictions. It quantifies the prop ortion of actual high-quality predictions—relative to the full range of high quality predictions made, which includes each genuine positives and false positives.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

Where:

True Positives (TP): The count of high-quality instances that the model appropriately anticipated.

Fake Positives (FP): The count number of instances that were incorrectly labelled as fantastic by using the model.

This metric becomes similar important in contexts where the consequences of false positives are significant, such as in medical diagnoses, where incorrectly diagnosing a healthy patient can lead to unnecessary anxiety and treatment, or in fraud detection, where falsely flagging legitimate transactions can harm customer trust and business operations.

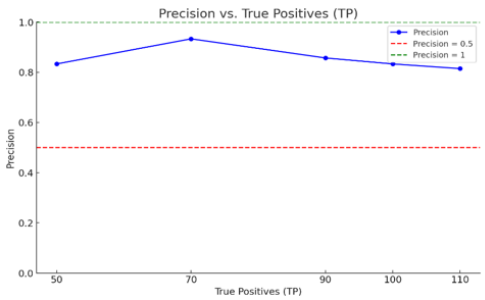

Fig. 3. Precision vs True Positives (TP)

By focusing on the quality of positive predictions, precision helps practitioners evaluate how reliably a model can identify true positive cases. This metric guides improvements in model performance and supports the decision-making process when the costs of errors need to be minimized.

E. Recall

Regularly called sensitivity or the proper advantageous rate, take into account is a important metric for assessing the performance of classification fashions, especially in scenarios in which datasets are imbalanced. It quantifies the version's capacity to correctly perceive effective instances among all actual positive instances. Do not forget is specially considerable in fields along with healthcare, in which failing to detect a fantastic case, like a disorder, may have extreme effects.

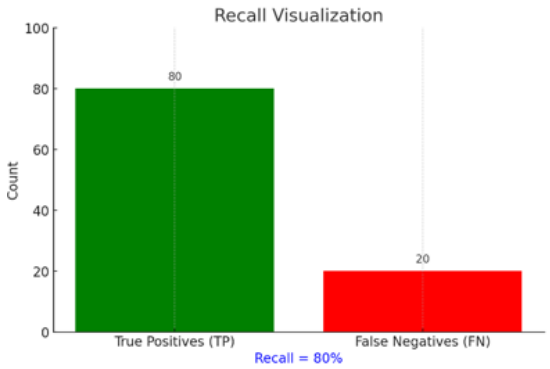

Fig. 4. Visualization of True Positives and False Negatives

In such contexts, an excessive don't forget rating shows that the model effectively captures most high quality times, minimizing the threat of false negatives.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

Components

- True Positives (TP): Times which are surely high quality and had been appropriately predicted as advantageous via the model.
- False Negatives (FN): Nice instances that the version did not identify effectively, misclassifying them as terrible.

By focusing on recall, practitioners can ensure that critical positive cases are not overlooked, even though this may lead to a higher number of false positives.

VI. CONCLUSION

In the end, the use of a Variational Autoencoder for flight fare prediction represents a vast improvement over conventional dynamic pricing models. VAEs are designed to analyze complex non-linear relationships within data, which is particularly useful in the highly variable and unpredictable airline market. In contrast to traditional models that often rely on simplistic assumptions about price elasticity and demand, VAEs can incorporate a much broader range of influencing factors, including seasonality, economic conditions, and consumer behavior. This allows for more nuanced fare predictions that adapt in real time to shifts in market dynamics, thereby improving overall forecasting accuracy.

Moreover, the VAE framework's ability to generate new data points based on learned patterns enables airlines to explore a wider set of pricing strategies. By capturing underlying patterns in historical airfare data, VAEs can identify potential opportunities for fare adjustments that conventional models might overlook. This flexibility not only leads to more accurate fare estimations but also helps airlines remain competitive in a crowded marketplace.

Incorporating Variational Autoencoders into flight fare prediction structures allows airlines to refine their pricing strategies, boost sales control, and supply tailor-made fare alternatives that enhance consumer pride. future improvements in making use of this for flight fare prediction could explore several critical instructions, which includes improving feature engineering by means of inclusive of numerous records assets like social media insights and monetary indicators to improve model precision. additionally, integrating this with reinforcement mastering may want to permit adaptive dynamic pricing strategies that respond to real-time market fluctuations.

REFERENCES

[1] E. Šimic and M. Begovic, ''Airport delay prediction using machine learning regression models as a tool for decision making process,'' in Proc. 45th Jubilee Int. Conv. Inf., Commun. Electron. Technol. (MIPRO), May 2022, pp. 841–846.

[2] S. S. B. T. Lincy, H. Al Ali, A. A. A. M. Majid, O. A. A. A. Alhammadi, A. M. Y. M. Aljassmy, and Z. Mukandavire, ''Analysis of flight delay data using different machine learning algorithms,'' in Proc. New Trends Civil Aviation (NTCA), Oct. 2022, pp. 57–62.

[3] D. Jadav, D. Patel, S. Thacker, A. Nair, R. Gupta, N. K. Jadav, and S. Tanwar, ‘‘EmReSys: AI-based efficient employee ranking and recommender system for organizations,’’ in Proc. Int. Conf. Comput., Commun., Intell. Syst. (ICCCIS), Nov. 2022, pp. 440–445.

[4] R. Vane, ‘‘Flight delay analysis and possible enhancements with big data,’’ Int. Res. J. Eng. Technol., vol. 3, no. 6, pp. 778–780, 2016.

[5] A. Dand, ‘‘Airline delay prediction using machine learning algorithms,’’ Ph.D. thesis, Wichita State Univ., College Eng., Dept. Ind., Syst. Manuf. Eng., Wichita, KS, USA, 2020.

[6] I.M. Almaameri and A. Mohammed, ‘‘Predicting airplane flight delays using neural networks,’’ in Proc. 5th Int. Conf. Eng. Technol. Appl. (IICETA), May 2022, pp. 579–584.

[7] T. Wang and S.-C. Chen, ‘‘Multi-task local–global graph network for flight delay prediction,’’ in Proc. IEEE 23rd Int. Conf. Inf. Reuse Integr. Data Sci. (IRI), Aug. 2022, pp. 49–54.

[8] C.-L. Wu and K. Law, ‘‘Modelling the delay propagation effects of multiple resource connections in an airline network using a Bayesian network model,’’ Transp. Res. E, Logistics Transp. Rev., vol. 122, pp. 62–77, Feb. 2019.

[9] Y. Wang, M. Z. Li, K. Gopalakrishnan, and T. Liu, ‘‘Timescales of delay propagation in airport networks,’’ Transp. Res. E, Logistics Transp. Rev., vol. 161, May 2022, Art. no. 102687.

[10] N. Chakrabarty, ‘‘A data mining approach to flight arrival delay prediction for American airlines,’’ in Proc. 9th Annu. Inf. Technol., Electromech. Eng. Microelectron. Conf. (IEMECON), Mar. 2019, pp. 102–107.

[11] M. F. Yazdi, S. R. Kamel, S. J. M. Chabok, and M. Kheirabadi, ‘‘Flight delay prediction based on deep learning and levenberg-marquart algorithm,’’ J. Big Data, vol. 7, no. 1, pp. 1–28, 2020.

[12] P. Meel, M. Singhal, M. Tanwar, and N. Saini, ‘‘Predicting flight delays with error calculation using machine learned classifiers,’’ in Proc. 7th Int. Conf. Signal Process. Integr. Netw. (SPIN), Feb. 2020, pp. 71–76.

[13] J.Yi, H. Zhang, H. Liu, G. Zhong, and G. Li, ‘‘Flight delay classification prediction based on stacking algorithm,’’ J. Adv. Transp., vol. 2021, pp. 1–10, Aug. 2021.

[14] Z. Shu, ‘‘Analysis of flight delay and cancellation prediction based on machine learning models,’’ in Proc. 3rd Int. Conf. Mach. Learn., Big Data Bus. Intell. (MLBDBI), Dec. 2021, pp. 260–267.

[15] R. Balamurugan, A. V. Maria, G. Baranidaran, L. MaryGladence, and S. Revathy, ‘‘Error calculation for prediction of flight delays using machine learning classifiers,’’ in Proc. 6th Int. Conf. Trends Electron. Informat. (ICOEI), Apr. 2022, pp. 1219–1225.

[16] Q. Li and R. Jing, ‘‘Flight delay prediction from spatial and temporal perspective,’’ Expert Syst. Appl., vol. 205, Nov. 2022, Art. no. 117662.

[17] K.Cai, Y. Li, Y.-P. Fang, and Y. Zhu, ‘‘A deep learning approach for flight delay prediction through time-evolving graphs,’’ IEEE Trans. Intell. Transp. Syst., vol. 23, no. 8, pp. 11397–11407, Aug. 2022.

[18] P.-J. Wen and C. Huang, ‘‘Machine learning and prediction of masked motors with different materials based on noise analysis,’’ IEEE Access, vol. 10, pp. 75708–75719, 2022.

[19] J. J. Dziak, D. L. Coffman, S. T. Lanza, R. Li, and L. S. Jermiin, ‘‘Sensitivity and specificity of information criteria,’’ Briefings Bioinf., vol. 21, no. 2, pp. 553–565, Mar. 2020.

[20] P. C. Emiliano, M. J. F. Vivanco, and F. S. de Menezes, ‘‘Information criteria: How do they behave in different models?’’ Comput. Statist. Data Anal., vol. 69, pp. 141–153, Jan. 2014.