IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 July 2023, accepted 24 July 2023, date of publication 26 July 2023, date of current version 3 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3298979

RESEARCH ARTICLE

Selection of Best Machine Learning Model to Predict Delay in Passenger Airlines

RAVI KOTHARI[1], RIYA KAKKAR[2], SMITA AGRAWAL[2], (Senior Member, IEEE), PARITA OZA[2], SUDEEP TANWAR[2], (Senior Member, IEEE), BHARAT JAYASWAL[3], RAVI SHARMA[4], GULSHAN SHARMA[5], AND PITSHOU N. BOKORO[5], (Senior Member, IEEE)

[1]Cognizant Technology Solutions, Bengaluru 560045, India
[2]Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat 382481, India
[3]IT School, Center of Excellence for Big Data, INS Valsura, Indian Navy, Jamnagar, Gujarat 361150, India
[4]Centre for Inter-Disciplinary Research and Innovation, University of Petroleum and Energy Studies, Dehradun 248001, India
[5]Department Electrical and Electronic Engineering Technology, University of Johannesburg, Johannesburg 2092, South Africa

Corresponding authors: Sudeep Tanwar (sudeep.tanwar@nirmauni.ac.in), Gulshan Sharma (gulshans@uj.ac.za), Smita Agrawal (smita.agrawal@nirmauni.ac.in), and Parita Oza (parita.prajapati@nirmauni.ac.in)

ABSTRACT Over the past years, flight delay has been a critical concern in the aviation sector due to the increased air traffic congestion worldwide. Moreover, it also prolongs the other flights, which can discourage users from traveling with the particular airline. As a result, we proposed a model to predict the overall flight delay using a random forest and path-finding algorithm. The proposed model focuses on searching flights (can be nonstop or connecting) between the source and destination at the earliest. The proposed model identifies the fastest flights between source and destination based on the input by the user using some open source/public Application Programming Interface (APIs), which are further inserted into Neo4j to convert it into a JavaScript Object Notation (JSON) format. Finally, the experimental results on the real-time data set show the proposed model's effectiveness compared to the state-of-the-art models. The results and analysis yield an accuracy of 98.2% for delay prediction on historical data using the random forest algorithm.

INDEX TERMS Flight delay, flight search, Neo4j, python, random forest.

I. INTRODUCTION

The aviation sector has fascinated people worldwide due to its various benefits of connecting communities, cultures, and businesses from different countries. Nowadays, users can directly book their flights in advance based on the destination with the help of advanced information and communication technologies compared to the traditional method of booking flights which involves reserving the flight manually. Furthermore, advanced technologies allow users to search for flights of varying prices that they can use to book flights based on their budget and affordability. However, the dynamic nature of aviation also suffers from the uncertainty of flight delay due to various factors such as a bad atmospheric environment, low visibility, or structural defects. The aforementioned challenges in the aviation sector cause air traffic congestion, leading to flight delays. With the rapid advancement in technology, flight detention seems to be one of the problems to circumscribe the development of the aviation industry and also the primary source for the airline passengers' dissatisfaction with the aviation service [1], [2], [3].

The main reasons for airline detainment include airline glitches, weather conditions, air traffic, etc., due to which users can get discouraged from booking a particular flight in the future. Further, based on the data report, due to weather conditions from November to March, the flight was found to be more delayed [4], [5]. Due to the aforementioned challenges related to flight delays, many researchers proposed their research work to improve aviation services for users to avoid flight delays. For example, the authors of [2] analyzed flight delay using various machine learning algorithms for other anomaly detection. But, accuracy using applied machine learning algorithms must be improved for early flight delay prediction. Next, Almaameri and Mohammed [6]

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Xiao.

TABLE 1. Abbreviations.

Abbreviation	Definition
NAS	National Airspace System
BTS	Bureau of Transportation Statistics
DOT	Department of Transportation
NOAA	National Oceanic and Atmospheric Administration
RNN	Recurrent Neural Network
MSE	Mean Square Error
SMOTE	Synthetic Minority Oversampling Technique
API	Application Programming Interface
JSON	JavaScript Object Notation
HTTP	Hypertext Transfer protocol

discussed flight delay prediction using neural networks in which the multilayer perceptron model is compared with the conventional classifiers for improved accuracy. Then, Wang and Chen [7] considered a multi-task learning model for flight prediction based on graph convolutional networks. Further, the authors of [8] considered a Bayesian network model to present a case study on the delay propagation effects of multiple resource connections in an airline network. They mainly focus to identify the weak links in a flight network based on past operational data. Next, Wang et al. [9] proposed a sliding correlation windows-based algorithm to amalyze the delay propagation in airport networks. However, the prediction model proposed by the above-mentioned researchers needs to improve their efficiency while predicting flight delay in terms of enhanced accuracy.

Therefore, considering an important aspect of accuracy for predicting flight delays, we proposed a model that can predict the flight delay using a random forest algorithm to warn users about the delay beforehand, and based on the user's feedback, airlines can improve their service to avail the better air traveling experience for the users. Moreover, the proposed model can search flights between source and destination without delay by considering the connecting or continuous flights. The proposed model utilizes Neo4j as a graph database which facilitates the data storage and processing of high-complexity data.

The paper is organized as follows: A related work in this domain is presented in section II. Section III discusses the proposed model, tools and technologies, and models for flight search and prediction. Detailed result analysis is presented in section IV. Section V elaborates on the discussion section. Finally, section VI presents the conclusion and future work, respectively. Further, Table 1 shows the abbreviations considered for the terms used in the proposed model.

II. RESEARCH CONTRIBUTIONS

- We have proposed a machine learning-based flight delay prediction model for passenger airlines using clustering sampling, random forest classifier and path-finding algorithm.
- We consider all parameters that affect delay(i.e., Air carrier, NAS, weather, late-arriving aircraft, and security).
- The proposed model utilizes Neo4j as a graph database which facilitates the data storage and processing of high-complexity data
- The model provides flight searching facilities between source and destination where it considers proposed delay prediction and shows the appropriate route through the Neo4J database.
- The proposed model is trained and tested on a real-time dataset and could result in 98.2% of overall accuracy.

III. RELATED WORK

Many researchers have incorporated machine learning and deep learning models to predict flight delays to avail efficient airline service to users. Some of the research works proposed by the researchers are: Chakrabarty [10] proposed a data mining approach for flight delay prediction to improve the performance of delay prediction in terms of accuracy. But, they should have focused on the rising security issues against malicious attacks while performing the flight delay prediction. Next, Yazdi et al. [11] utilized the deep learning and levenberg-marquart algorithm for flight delay prediction. However, there is no discussion about delay prediction in the real-time scenario. Then, Yi et al. [13] considered a stacking algorithm to predict and classify the flight delay to improve the overall stability of the proposed system. Later, Balamurugan et al. [15] applied the error calculation mechanism for predicting flight delays using a machine learning classifier. Later, Cai et al. [17] utilized the time-evolving graphs to perform flight delay prediction using a deep learning approach. Finally, the aforementioned prediction models presented by the researchers focused on flight delay prediction so that users can book their flights in advance based on the prediction. Nevertheless, the proposed prediction models do not efficiently predict flight delay with improved accuracy in the real-time scenario. Moreover, they should have explored various path-finding algorithms to avail less expensive flights for users. Thus, based on the aforementioned challenges of the state-of-the-art prediction models, we proposed a model to predict flight delay using a random forest and path-finding algorithm that facilitates users to reserve available flights with improved accuracy and efficiency. Table 2 shows the comparative analysis of the state-of-the-art prediction models with the proposed model highlighting highlights the identified research gaps in the related research works raising raises the need for the proposed prediction model using a random forest algorithm for flight delay prediction.

IV. PROPOSED MODEL

The proposed model works to find flights between the source and destination location based on the traveling requirement of the users in minimum time while considering nonstop as well as connecting flights. In addition, the proposed model also predicts the delay of the flights based on parameters such as air carrier, National Airspace System (NAS), weather,

TABLE 2. Comparative analysis of state-of-the-art flight delay predicting models with the proposed model.

Author	Year	Objective	Pros	Cons
Chakrabarty *et al.* [10]	2019	Proposed a data mining approach for flight delay prediction	Improved performance and accuracy	Security issues against malicious attacks are not considered
Yazdi *et al.* [11]	2020	Disussed the flight delay prediction with deep learning and levenberg-marquart algorithm	Better accuracy	Real-time scenario needs to be analysed
Meel *et al.* [12]	2020	Performed flight delay prediction utilizing machine learning classifiers	Yields best value for delay prediction	Performance can be improved in terms of accuracy based on the processing power
Yi *et al.* [13]	2021	Utilized stacking algorithm for flight delay classification and prediction	Improved stability	Need to focus on other important aspect for delay prediction
Shu [14]	2021	Performed flight delay and cancellation prediction using machine learning models	Better performance for validation set	Needs to improve model accuracy in real-time scenario
Balamurugan *et al.* [15]	2022	Considered machine learning classifier for predicting flight delays along with the error calculation	Better accuracy	Need to consider the model in the real-time scenario
Li *et al.* [16]	2022	Proposed a framework for predicting flight delay considering spatial and temporal perspective	Real-time monitoring for improved accuracy prediction	No consideration of important factors such as weather conditions while performing prediction
Cai *et al.* [17]	2022	Presented a deep learning approach for flight delay prediction utilizing the time-evolving graphs	Low execution time	Security issues against data manipulation, spoofing, and data modification attacks
The proposed model	2023	Proposed a flight delay prediction model using random forest and path finding algorithm	Better accuracy and high efficiency	-

late-arriving aircraft, and Security. We applied a random forest algorithm for delay prediction. This section discusses the tools and technology used for the proposed work and the models for flight search and flight delay prediction. Figure 1 shows the overall proposed model to perform flight delay prediction using a random forest algorithm and to search the affordable flights for users using the path-finding algorithm.

A. TOOLS AND TECHNOLOGIES

This section represents tools and technologies used to build the proposed model by specifying a detailed and comprehensive description of the tools and technologies. For example, Table 3 presents a detailed explanation of tools and technologies such as Neo4j, python, tableau, Kiwi Application Programming Interface (API), etc. Further, before discussing the flight delay prediction using the random forest algorithm, we need to discuss the model for searching flights using the path-finding algorithm.

B. MODEL FOR SEARCHING FLIGHTS

This section shows the methodology adopted to build a model for the flight search. Then, we presented the data model for the flight search and execution flow of the model for searching flights.

1) DATA MODEL

Aviation industry networks are one of the most complex and scale-free networks. This network has highly connected data with highly complex inter-entity relationships, so a graph database can be recommended for such a complex network. So, to store and process such highly connected data, we utilize a graph database and create a model for searching the flights. A graph database is a database that is designed for viewing data connections as equally relevant to the data itself. It pursues to hold data without restricting it to a predefined standard. A graph includes two parts: a node and a relationship. Every node portrays an entity (a flight, a person, a place, a thing, a category, or another piece of data), and each relationship portrays how two nodes are related. Figure 2 represents the design based on the model given by Max De Marzi. The problem with this model is the association of the data. If we use the date as a relationship, then there would be one extra traversal in the execution of each query from Airport to AirportDay. Instead, we have used the AirportDay node as the airport + date node (i.e., Airport code + date in milliseconds) to make the execution faster but decrease the user experience as we can't calculate milliseconds to the exact date without a converter. So, we have created a month and a day node in addition to the AirportDay node.

2) FLOW OF EXECUTION

The flow of the execution of the proposed model is presented in Figure 3. End-user (E_u is asked to enter source (α_{E_u}) and destination addresses (β_{E_u}) for searching the appropriate flights. These addresses are passed to Here API and output latitude and longitude of both the addresses, i.e., ($\epsilon^{\alpha_{E_u}}$, $\varepsilon^{\alpha_{E_u}}$)

TABLE 3. Tools and technologies used for implemamtation of the proposed model.

Tool/Technologies	Description
Neo4j	• Neo4j is a NoSQL database classified beneath neath graph databases, and it follows the mathematical concept of trees. • A graph includes two parts: a node and a relationship. Every node portrays an entity (a person, place, thing, category or different piece of data), and every relationship portrays how nodes are related. • Neo4j is extensively utilized in social media networks to link one or more people together.
Python	• Python is a programming language that enables us to operate faster and to integrate programs easily.
Tableau	• A tableau is software that can be used for data visualization. We used it to visualize all the airports with their connections in the world.
Py2neo	• Py2neo is a library for interacting with Neo4j from within Python. It supports protocols like HTTP and Bolt.
Kiwi API	• Kiwi API is for providing fare aggregator, all-in-one search engine and airplane ticket booking, and overland transportation. • Searching can be done using simple parameters such as places and flight API endpoints. • Proposed model uses python to get data from the Kiwi API and insert it into Neo4j, a combination of nodes and relationships.
Here API	• The "Here" Routing API is an HTTP JSON REST API it enables to calculate routes among one or more locations inclusive of the traffic delay. • "Here" API's Geocoding function is also used to find the coordinates of the origin and the destination addresses for further use.
Lufthansa Public API	• Lufthansa's Public API gives a way to find the airport from the coordinates acquired from the Here Geocoder.

and $(\epsilon^{\beta_{E_u}}, \varepsilon^{\beta_{E_u}})$. Lufthansa API is fed with the latitude and longitude of both addresses to get the airport code, latitude, and longitude of the nearest airport from source and destination addresses. Airport-code, latitude, and longitude information is provided to HERE routing API, which determines the travel time from (source, destination) address to (source, destination) airport. The associations mentioned above are represented as follows:

$$E_u = \alpha_{E_u}, \beta_{E_u} \tag{1}$$

$$\{\alpha_{E_u}, \beta_{E_u}\} \xrightarrow{HereAPI} \{(\epsilon^{\alpha_{E_u}}, \varepsilon^{\alpha_{E_u}}), \times (\epsilon^{\beta_{E_u}}, \varepsilon^{\beta_{E_u}})\} \tag{2}$$

$$\{(\epsilon^{\alpha_{E_u}}, \varepsilon^{\alpha_{E_u}}), (\epsilon^{\beta_{E_u}}, \varepsilon^{\beta_{E_u}}\} \xrightarrow{input} LUFTHANSAAPI \tag{3}$$

Furthermore, KIWI API sends the flight data to Neo4j via the Py2Neo database driver. Neo4j processes and provides flight search visualization on the Neo4j browser, which is explained in Algorithm 1.

Algorithm 1 Algorithm for Flight Search

Input: $\alpha_{E_u}, \beta_{E_u}, \epsilon^{\alpha_{E_u}}, \varepsilon^{\alpha_{E_u}}$
Output: Flight search data

procedure Flight_search($\alpha_{E_u}, \beta_{E_u}$)
    Input source and destination address from user to forward it to HERE Geocoding API
    $user \xrightarrow{\{\alpha_{E_u}, \beta_{E_u}\}} HEREAPI$
    $\{\alpha_{E_u}, \beta_{E_u}\} \xrightarrow{HereAPI} \{(\epsilon^{\alpha_{E_u}}, \varepsilon^{\alpha_{E_u}}), (\epsilon^{\beta_{E_u}}, \varepsilon^{\beta_{E_u}}\}$
    $\{(\epsilon^{\alpha_{E_u}}, \varepsilon^{\alpha_{E_u}}), (\epsilon^{\beta_{E_u}}, \varepsilon^{\beta_{E_u}}\} \xrightarrow{input}$ Lufthansa API
    Lufthansa API yields the airport-code, latitude, and longitude of the nearest airport from source and destination address.
    Send the output of Lufthansa API to HERE ROUTING API
    HERE Routing API yields traveling time from (source, destination) to airport
    KIWI API fowards flight data to Neo4j browser
end procedure

C. MODEL FOR DELAY PREDICTION

According to airline industries, many reasons for flight detainment include airline glitches, weather conditions, sustentation problems with the aircraft, air traffic jams, late advent of the aircraft to be used for the flight from a former flight, and security issues [4]. For the same source of data, data cleaning, delay calculation, and delay prediction through random forest algorithm are discussed in this section.

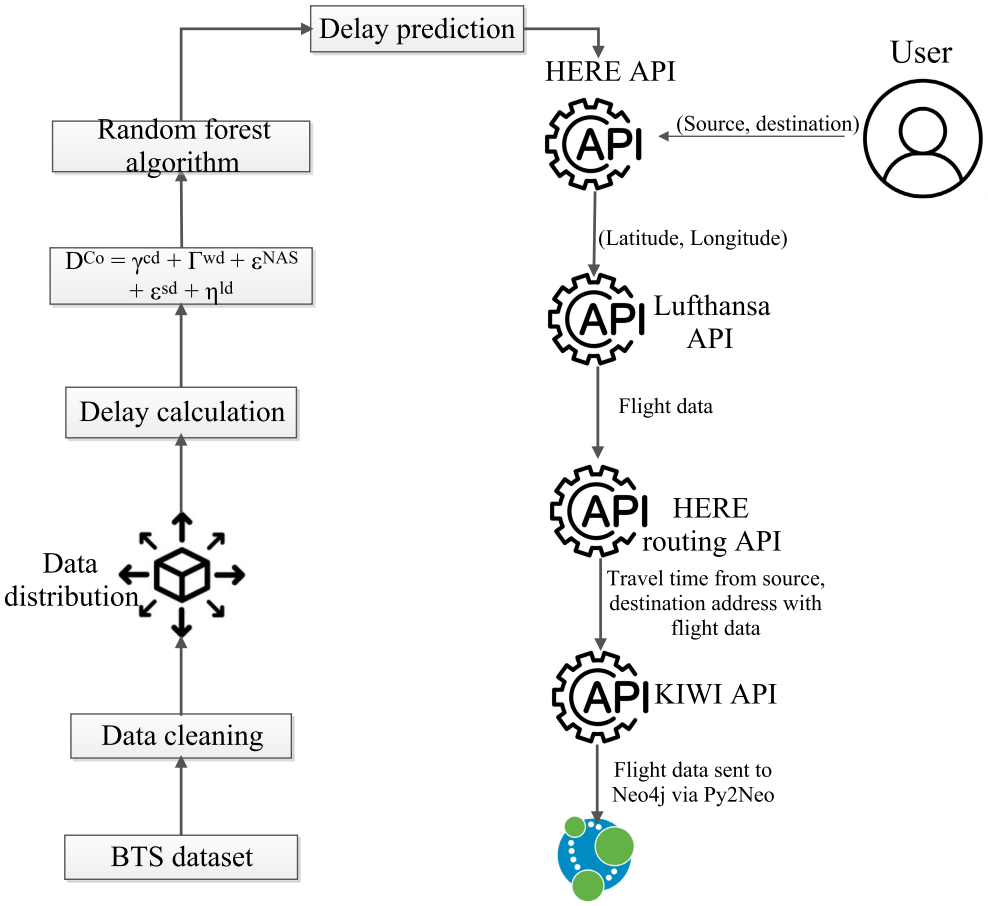

FIGURE 1. The proposed model.

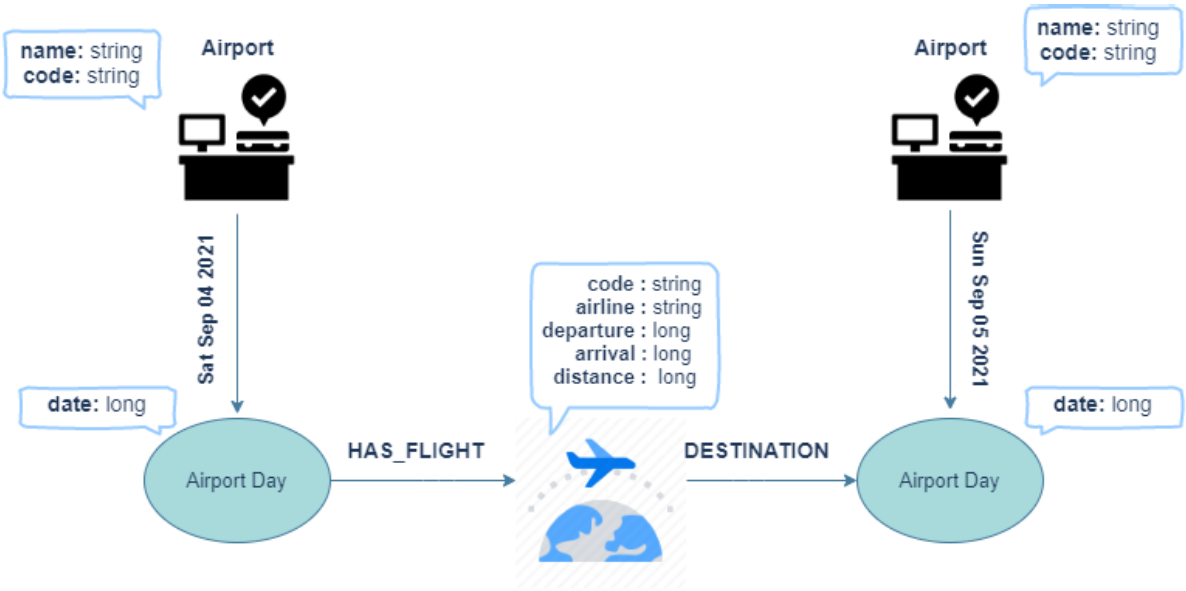

FIGURE 2. Design based on model given by max De marzi.

1) DATA SOURCES

- The Bureau of Transportation Statistics (BTS) of the US Department of Transportation (DOT) offers accurate and credible records on the US transportation system. It facilitates data of all commercial flights in the USA in conjunction with the reason for flight delay.

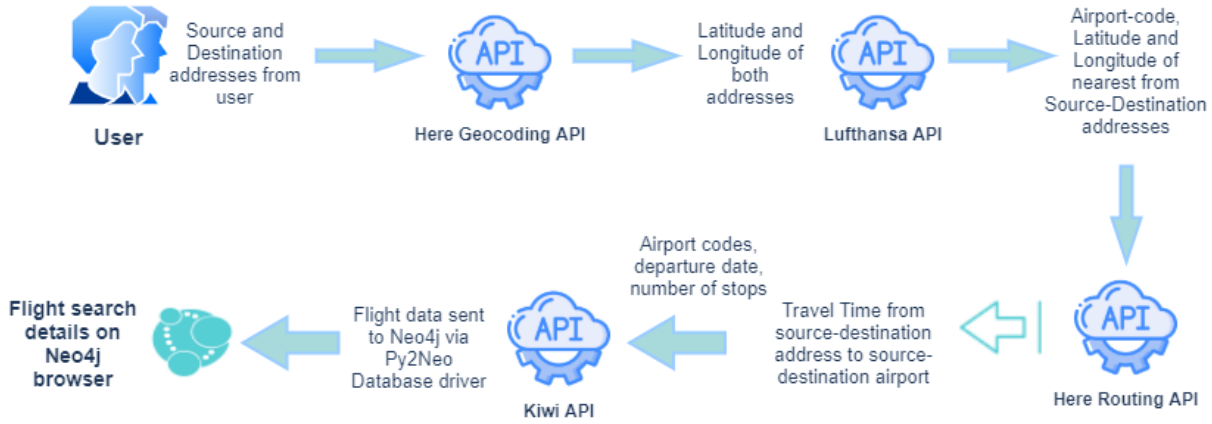

FIGURE 3. Flight search program' flow of execution.

- The US National Oceanic and Atmospheric Administration (NOAA) is an organization that is a part of the US Department of Commerce. NOAA's mandate is involved with the oceans and the atmosphere. It provides departure airport weather details, including temperature, humidity, air pressure, and rainfall type.

1) **Data Cleaning**
   - Only the data fields like the year, month, date, day of the week, departure and Arrival time (local), and arrival delay indicator (0 if the delay time is less than 15 minutes, one if the delay time is larger than or equal to 15 minutes) affects the flight delays from BTS dataset.
   - We used data for August because June to August is the dry season in the US, leading to unusually long delays due to the thunderstorms, snow, and rainfall that don't skew our data. Based on the visualization of Figure 4, most of the area received near or below-average rainfall.

2) **Data attributed used for processing**
   - Columns: CRS departure-elapsed time, origin, destination, distance, carrier mean distance, and origin delay.
   - Considering 80:20 ratio of data splitting, 2,40,000 rows for training and 60,000 rows for testing are used out of total 3,00,000 rows/flights.

3) **Delay Calculation**
   - Total delay computation for overall flight delay prediction can be performed considering the parameters carrier delay (γ^{cd}), weather delay (Γ^{wd}), NAS delay (ϵ^{NAS}), security delay (ε^{sd}), and late aircraft delay (η^{ld}). Total delay computation (D^{Co}) is expressed as follows:

$$D^{Co} = \gamma^{cd} + \Gamma^{wd} + \epsilon^{NAS} + \varepsilon^{sd} + \eta^{ld} \quad (4)$$

2) RANDOM FOREST ALGORITHM

The main motive behind adopting a random forest algorithm for flight delay prediction is an easy yet strong one, i.e., the collective opinion. The cause of the random forest model working efficiently is a lot of relatively disconnected models (trees) working as a single force can outmatch any individual models. Random forest uses ensemble learning, which uses multiple algorithms to get better prediction performance from any constituent algorithms alone [10]. Prediction in the random forest algorithm is analyzed based on the majority of the class or by considering the regression tree average.

- By sampling N aimlessly if the calculation of cases in the training set is N but with a substitute from the original data. Then, the sample tends to be a training set for the growing tree.
- For K input variables, the variable k is selected so that kK can be specified at each node, k variables are selected randomly from the K, and the best split on the k is used for splitting the node. During the forest growing, the value of k is kept constant.
- Here, each tree is grown to the largest feasible extent without any usage of pruning.

3) MODEL SELECTION CRITERIA: RANDOM FOREST

The main reason for considering the random forest algorithm for predicting flight delay can be explained with the information criteria, which works on the principle of probabilistic measure to select the best model among various machine learning models by calculating the values of various selection measures such as Akaike information criterion (AIC),Schwartz Bayesian information criterion (SBIC), and Hannan-Quinn information criterion (HQIC) [18]. Now, information criteria, i.e., AIC, SBIC, and HQIC, measures the relative information loss which is considered to select the best model based on their calculated values for the considered machine learning models. The model which yields the

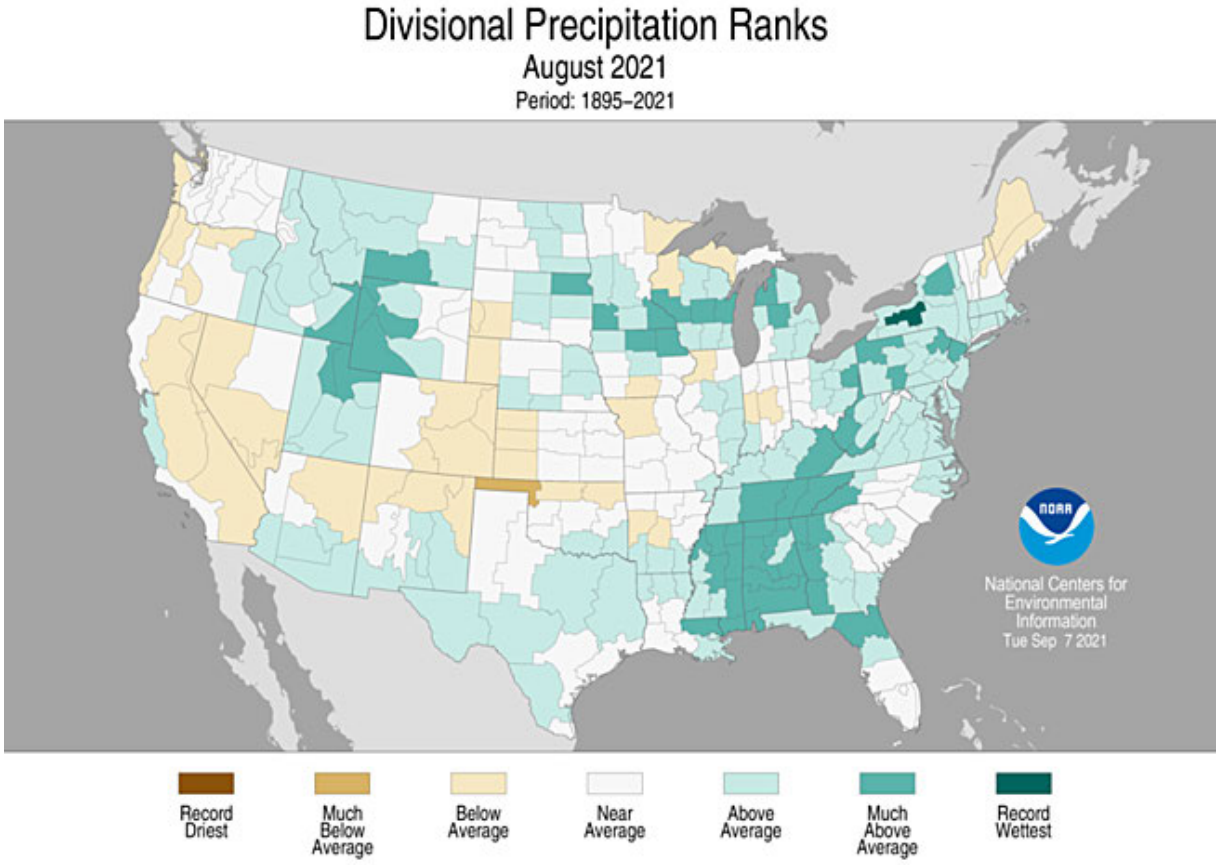

FIGURE 4. Divisional Precipitation Ranks (Rainfall) for August 2021.

minimized information loss for information criteria can be selected for predicting the flight delay accurately [19], [20]. Furthermore, we have utilized various machine learning models such as gradient boosting and decision tree to compare them with the proposed random forest algorithm for predicting the flight delay. Moreover, we have performed the statistical analysis of the proposed model considering the information criteria, i.e., AIC, SBIC, and HQIC, as a information loss to select the best machine learning model for prediction. Table 4 compares various machine learning models, such as gradient boosting classifier and decision tree, with the random forest algorithm associated with their calculated values for predicting flight delay for passenger airlines. Thus, random forest as a minimum information loss, i.e., value of 8.551, 14.376, and 11.358 for information criterion AIC, SBIC, and HQIC proves to be the best machine learning model for predicting flight delay than the other machine learning models, which also proves the better parsimony of the proposed model. The minimum information loss of the proposed random forest model reflects the fewer parameters or predicting variables while performing the flight delay prediction, further yielding the optimal parsimony for the proposed model considering the information criterion such as AIC, SBIC, and HQIC.

TABLE 4. Comparison of various machine learning model with the proposed model based on the information criterion.

Machine learning model	AIC	SBIC	HQIC
Gradient Boosting Classifier	10.631	16.729	13.289
Decision Tree	11.923	18.124	15.899
Proposed model with random forest	8.551	14.376	11.358

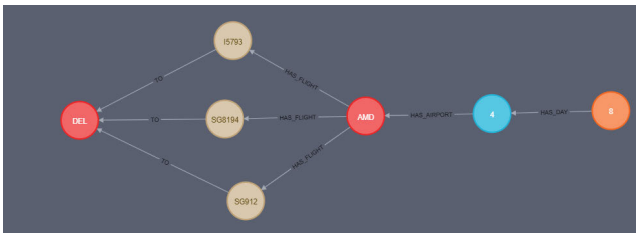

FIGURE 5. Execution of query in Neo4j Browser.

V. RESULT ANALYSIS

1) FLIGHT SEARCH

We have searched flights between Ahmedabad and Delhi from 17th May 2020 to 25th August 2020, i.e., for the next 100 days. Further, the data we got from Here API is inserted into Neo4j. In Figure 5, we focus on searching the flights from

Algorithm 2 Algorithm for Flight Delay Prediction

Input: γ^{cd}, Γ^{wd}, ϵ^{NAS}, ε^{sd}, η^{ld}
Output: Accuracy

1: **procedure** Delay_prediction(γ^{cd}, Γ^{wd}, ϵ^{NAS}, ε^{sd}, η^{ld})
2: Apply data cleaning on the considered BTS dataset
3: Perform data pre-processing on the considered parameters.
4: $D^{Co} = \gamma^{cd} + \Gamma^{wd} + \epsilon^{NAS} + \varepsilon^{sd} + \eta^{ld}$
5: Random forest is applied to train model in 80:20 ratio.
6: Flight delay is predicted with 98.2% accuracy.
7: **end procedure**

Ahmedabad to Delhi from 17th March 2020 to 25 August 2020, i.e., 100 days in which the data returned by the API is an adequately formed JavaScript Object Notation (JSON).

2) DELAY PREDICTION

We used 3,00,000 rows to apply the random forest model on the considered BTS dataset. We used cluster sampling for delay prediction due to the huge data. Cluster sampling also requires fewer resources than other forms of sampling. We used 20% (60,000) rows for testing and 80% (2,40,000) rows for training the model. We found that CRS departure time and scheduled departure time majorly impact the overall delay, followed by CRS elapsed time and departure delay in minutes. We also predicted delay using custom data, which returned the delay probability in percentage. A probability between 10 and 25 percent would mean that the flight would reach the destination on time, even if the flight takes off an hour late because airlines increase their speed mid-air to be on time. On the other hand, a probability of above 25 % would mean that the flight will get delayed. And a probability above 70% would mean that the flight will likely be canceled based on the unusual delay.

Example :

- When a flight carrier with a mean flying distance of 488 kilometers is going for a flight of 670 kilometers within its capability, the flight is likely to get on time. Also, if the flight is at night (20:49 hours), then even after a 51-minute departure delay, the airline can increase its speed mid-air to match the scheduled time.
- When a flight carrier with a mean distance of 455.87 kilometers is going on a flight of 2000 kilometers it is in the middle of the day (16:00 hours) with a departure delay of 116 minutes, then the flight is more likely to get canceled or delayed.

3) RANDOM FOREST ALGORITHM RESULTS

The random forest algorithm has been applied to the considered BTS dataset consisting of 3,00,000 rows to further perform the cluster sampling for flight delay prediction. The flight delay prediction is performed on the BTS dataset associated with the features or parameters, i.e., γ^{cd}, Γ^{wd}, ϵ^{NAS}, ε^{sd}, and η^{ld} by classifying it into 80% training data and 20% testing data. Moreover, the random forest algorithm is applied to the dataset along with ensemble learning for regression that uses multiple machine learning models for better prediction [21]. For that, mean square error (MSE) can be computed to predict flight delay using a random forest algorithm, which is expressed as follows:

$$MSE = \frac{\sum_{k=1}^{N}(M_k - A_k)^2}{N} \tag{5}$$

where N is the parameter value for the prediction and training dataset. M_k denotes the value obtained based on the applied model, and A_k signifies the actual value while predicting the training dataset. Furthermore, Figure 6 shows the confusion matrix for data which decides whether the obtained predicted output is correct.

	Positive	**Negative**	
Positive	58848	71	**Sensitivity**
			0.998
Negative	984	97	**Specificity**
			0.089
	Precision	**F1-Measure**	**Accuracy**
	0.983	0.155	0.982

FIGURE 6. Confusion matrix.

Moreover, in Figure 7, we have performed and analyzed the accuracy of the proposed flight delay prediction model with the conventional approaches to highlight the improved performance of the proposed model. It can be observed from the graph that the proposed model predicts flight delay prediction with an accuracy of 98.2%, which proves to be an effective model for the users and outperforms conventional approaches in terms of efficiency and accuracy.

Algorithm 2 highlights the overall flight delay prediction with an accuracy of 98.2% by performing the data pre-processing and cluster sampling on the considered dataset utilizing the parameters. Further, a random forest algorithm is applied to predict the overall flight delay prediction to improve airline service for users.

VI. DISCUSSION

In this section, a comparative study of all the conventional approaches proposed by the researchers is analyzed with the proposed model based on the parameters such as delay reasons, sampling technique, algorithm, data source, and accuracy. Table 5 shows the comparative analysis of the conventional flight delay prediction approaches with the proposed model. For example, Choi et al. [22] predicted delay caused by climate using BTS and NOAA (weather) datasets. Further, they have considered wind indicators, visibility, precipitation (water and snow), and combination indicator code from the NOAA dataset. Then, Prabakaran et al. used

TABLE 5. Comparative analysis of conventional approaches with the proposed model based on the considered parameters.

Author	Delay Reason	Sampling Technique	Algorithm(s)	Data Source	Accuracy
[22]	Weather	SMOTE	Decision tree, random forest, AdaBoost, K-Nearest Neighbours	BTS, NOAA	80.36%
[23]	All*	Cluster	Regression	BTS	4.81%
[10]	Late arriving aircrafts	Random-SMOTE	Gradient Boosting Classifier	BTS	85.73%
[24]	All*	SMOTE, Tomek	Gradient Boosting, random forest, extra trees, AdaBoost	BTS, World weather online API	94.35% (Arrival)
[25]	All*	Simple Random	Gradient Boosted Decision Tree	BTS	Arrival: 92.31%, Departure: 94.85%
[26]	All*	Simple Random	Naïve Bayes, Bayesian Network	BTS	75%
Proposed model	All*	Cluster sampling	Random forest	BTS, NOAA	98.2%

*All** : Air carrier, NAS, weather, late-arriving aircraft, and security

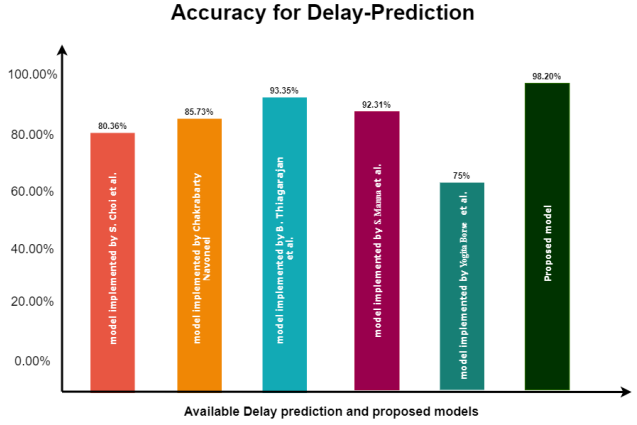

FIGURE 7. Accuracy for delay prediction.

various parameters of date like a month, year, day, and week along with origin and destination airport, schedule arrival-departure, time, delay, and distance from the BTS dataset for performing prediction of the flight delay. Moreover, they have proposed two models in which model 1 achieved 10.2 minutes of difference between predicted and actual delays. In model 2, 7.7 minutes of difference between forecasted & actual delays was achieved.

Thiagarajan et al. [24] utilized 12 airlines' on-time performance data attributes and 24 weather attributes from the BTS dataset and World Weather Online API, respectively. They developed a two-stage predictive model, where they predict the incidence of a delay and then the delay in minutes. This model was used to forecast both delays, i.e., arrival and departure delays. Manna et al. [25] considered a day of the week, carrier, origin-destination, airport IDs, CRS time, and delay from the BTS dataset to implement flight delay prediction. They have obtained an arrival accuracy of 92.31% and a departure accuracy of 94.85%. Furthermore, Borse et al. [26] utilized the simple random sampling technique along with the Naive Bayes and Bayesian network algorithm to yield an accuracy of 75% in a flight delay prediction system. Thus, the proposed model for flight delay prediction exhibits improved, and better accuracy of 98.2% predicted with cluster sampling and random forest algorithm than the conventional approaches.

VII. CONCLUSION

In this paper, we proposed a random forest and path-finding prediction model to ease the flight searching task in which KIWI API forwards the flight data to Neo4j and Py2Neo database drivers. The proposed model provides nonstop and multi-stop flight information. Moreover, the random forest algorithm predicts the overall flight delay for a better user's air traveling experience. Experimental results show that the proposed work outperforms other related work of the domain and achieves better accuracy (98.2%) using a random forest algorithm. In the future, we will predict the delay with higher accuracy using Graph Recurrent Neural Network (RNN) due to the weather and other essential aspects by considering the relevant parameter in the dataset. We will add an option to book the flights as well, and RNN will convert the data into a sequence of nodes, but that will make it lose the data structure, so we need to develop a Graph RNN to maintain the structure of the data.

ACKNOWLEDGMENT

The numbers specify the sequence of the authors along with their contributions:
1 (writing- original draft, review, and editing), 2 (writing-original draft, methodology, figures), 3 (conceptualization, methodology, and writing-original draft), 4 (conceptualization, methodology, and writing-review and editing), 5 (conceptualization, methodology, writing-original draft, and visualization), 6 (methodology, investigation, and visualization), 7 (conceptualization, methodology, writing-review and editing), 8 (conceptualization, visualization, and investigation), 9 (methodology, visualization, writing- review and editing).

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.
[21] K. Patel, D. Mehta, C. Mistry, R. Gupta, S. Tanwar, N. Kumar, and M. Alazab, "Facial sentiment analysis using AI techniques: State-of-the-art, taxonomies, and challenges," *IEEE Access*, vol. 8, pp. 90495–90519, 2020.
[22] S. Choi, Y. J. Kim, S. Briceno, and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," in *Proc. IEEE/AIAA 35th Digit. Avionics Syst. Conf. (DASC)*, Sep. 2016, pp. 1–6.
[23] N. Prabakaran and R. Kannadasan, "Airline delay predictions using supervised machine learning," *Int. J. Pure Appl. Math.*, vol. 119, no. 7, pp. 329–337, 2018.
[24] B. Thiagarajan, L. Srinivasan, A. V. Sharma, D. Sreekanthan, and V. Vijayaraghavan, "A machine learning approach for prediction of on-time performance of flights," in *Proc. IEEE/AIAA 36th Digit. Avionics Syst. Conf. (DASC)*, Sep. 2017, pp. 1–6.
[25] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta, and S. Barman, "A statistical approach to predict flight delay using gradient boosted decision tree," in *Proc. Int. Conf. Comput. Intell. Data Science(ICCIDS)*, Jun. 2017, pp. 1–5.
[26] Y. Borse, D. Jain, S. Sharma, and A. Z. Viral Vora, "Flight delay prediction system," *Int. J. Eng. Res.*, vol. V9, no. 3, pp. 88–92, Mar. 2020.

RAVI KOTHARI received the MCA degree from the Institute of Technology, Nirma University, in 2021. He is currently a Programmer Analyst Trainee with Cognizant Technology Solutions. He is also an American Client Specialising in virtualization. Helping them automate blockchain system testing with Data Driven Model Based Testing (DDMBT) on various environments (on-prem, AWS, and Azure) using Python.

RIYA KAKKAR received the bachelor's and M.Tech. degrees from Banasthali Vidyapith, Jaipur, India, in 2018 and 2021, respectively. She is currently a full-time Ph.D. Research Scholar with the Computer Science and Engineering Department, Nirma University, Ahmedabad, Gujarat, India. She is an active member of the ST Research Laboratory. She has authored or coauthored some publications, (including papers in SCI Indexed Journal and IEEE ComSoc sponsored International Conference). Some of her research findings are published in top-cited journals and conferences, such as IEEE Systems Journal, IEEE Internet of Things Journal, *Journal of Information Security and Applications*, *International Journal of Energy Research* (Wiley), IEEE CITS, IEEE ICC, and IEEE INFOCOM. Her research interests include electric vehicles, blockchain technology, 5G communication networks, and machine learning.

SMITA AGRAWAL (Senior Member, IEEE) received the MCA degree from Gujarat Vidhyapith, in 2004, and the Ph.D. degree in big data analytics from CHARUSAT University, in 2019. She works in the area of big data analytics, parallel processing, web development, and the IoT. She has been an Assistant Professor with the Computer Science and Engineering Department, since 2009. She has a teaching experience of more than 14 years. She has conducted ISTE approved short term training program in the field of web services using PHP. She is involved in teaching courses at both undergraduate and postgraduate level. She has published several Scopus/SCIE indexed research papers in national and international conferences and journals. She is a member of IEEE, CSI, ACM, and ISTE. She serves as a member of the Program Committees and the Session Chair for International Conferences. She serves as a reviewer for indexed international journals.

PARITA OZA received the M.Tech. degree in information and communication technology from Nirma University. She is currently pursuing the Ph.D. degree with Pandit Deendayal Energy University, Gandhinagar, India. She is an Assistant Professor with the Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad. She is involved in teaching courses at both undergraduate and postgraduate levels. She has published several research papers in national and international conferences and journals. She has mentored many B.Tech. and M.Tech. Projects. Her research interests include image processing, computer vision, and medical imaging. She serves as a member of the Program Committees and the Session Chair for International Conferences. She serves as a reviewer for indexed international journals.

SUDEEP TANWAR (Senior Member, IEEE) received the B.Tech. degree from Kurukshetra University, India, in 2002, the M.Tech. degree (Hons.) from Guru Gobind Singh Indraprastha University, Delhi, India, in 2009, and the Ph.D. degree with specialization in wireless sensor network, in 2016. He is currently a Full Professor with the Computer Science and Engineering Department, Institute of Technology, Nirma University, India. He is also a Visiting Professor with Jan Wyzykowski University, Polkowice, Poland, and the University of Pitesti, Pitesti, Romania. He is leading the ST Research Laboratory, where group members are working on the latest cutting-edge technologies. He has authored seven books and edited 22 books and more than 350 technical articles, including top journals and top conferences, such as IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE WIRELESS COMMUNICATIONS, IEEE NETWORKS, IEEE ICC, IEEE GLOBECOM, and IEEE INFOCOM. He initiated the research field of blockchain technology adoption in various verticals, in 2017. His H-index is 67. He actively serves his research communities in various roles. His research interests include blockchain technology, wireless sensor networks, fog computing, smart grids, and the IoT. He is a Final Voting Member of the IEEE ComSoc Tactile Internet Committee, in 2020. He is a member of CSI, IAENG, ISTE, and CSTA, and a member of the Technical Committee on Tactile Internet of IEEE Communication Society. He was awarded the Best Research Paper Awards from the IEEE IWCMC-2021, the IEEE GLOBECOM 2018, the IEEE ICC 2019, and the Springer ICRIC-2019. He has served as a member of the Organizing Committee for many international conferences, such as the Publication Chair for the FTNCT-2020, the ICCIC 2020, and the WiMob2019, a member of the Advisory Board for the ICACCT-2021 and the ICACI 2020, the Workshop Co-Chair for CIS 2021, and the General Chair for the IC4S 2019 and 2020, and the ICCSDF 2020. He is also serving on the Editorial Board of *Computer Communications*, *International Journal of Communication System*, and *Security and Privacy*.

BHARAT JAYASWAL served as a Commander with Indian Navy, with having previous experiences, such as an Electrical Officer in Missile and Weapon Systems, Onboard Capital Warships, the Manager of Missile Systems, a IT Officer onboard ships/units. He is currently appointed as the Officer-in-Charge, IT School, INS Valsura, and the Center of Excellence for AI/Big Data. Some of his projects include Maritime Anomaly Detection and steered as the Head of the Center of Excellence for AI/Big Data Analysis.

RAVI SHARMA is currently a Professor with the Centre for Inter-Disciplinary Research and Innovation, University of Petroleum and Energy Studies, Dehradun, India. He is passionate in the field of business analytics and worked in various MNCs as the Leader of various software development groups. He has contributed various articles in the area of business analytics, prototype building for startup, and artificial intelligence. He is leading academic institutions, as a Consultant, to uplift research activities in inter-disciplinary domains.

GULSHAN SHARMA received the B.Tech., M.Tech., and Ph.D. degrees. He was a Postdoctoral Research Fellow at Faculty of EBIT, University of Pretoria, South Africa, from 2015 to 2016. He is a Y Rated Researcher from National Research Foundation (NRF) of South Africa. He is working as a Academic Editor of International Transactions on Electrical Energy System Journal and Journal of Electrical and Computer Engineering, Hindawi. He is presently working as Senior Lecturer in the Faculty of Engineering and the Built Environment of the University of Johannesburg, South Africa. He has published more than 100 research papers in international journals and conferences and he has been continuously engaged in guiding research activities at graduate/post-graduate and Ph.D. levels. His research interests include power system operation and control, renewable power generation, FACTS and application of AI techniques to power systems.

PITSHOU N. BOKORO (Senior Member, IEEE) received the bachelor's degree in electrical engineering from the Durban University of Technology, the master's degree in electrical engineering from the University of Johannesburg, and the Ph.D. degree from the University of the Witwatersrand, Johannesburg. He is an Associate Professor with the Faculty of Engineering and the Built Environment of the University of Johannesburg, South Africa. He has published over 100 research papers (which include over 50 journal articles and 70 conference papers) in indexed journals and peer reviewed conference proceedings. He authored a couple of book chapters in reputed books published by IGI-Global and IET. His major research interests include renewable energy systems, power systems, power system reliability, distributed generation, surge arresters, insulation and dielectrics, power quality, condition monitoring, microgrid, the Internet of Things and applied artificial intelligence. He holds Senior Membership with the South African Institute of Electrical Engineers (SMSAIEE) as well as with the Institute of Electrical and Electronics Engineers (SMIEEE). He serves as a specialist editor in *Energy and Power Systems for the SAIEE Africa Research Journal* (ARJ). He has supervised to completion over 18 postgraduate students (which include maste's and doctoral students).