Journal of Techniques, ISSN: 2708-8383, Vol. 6, No. 3, 2024, Pages 26-34
DOI: https://doi.org/10.51173/jt.v6i3.2481

JOURNAL OF TECHNIQUES

Journal homepage: http://journal.mtu.edu.iq

RESEARCH ARTICLE - COMPUTER ENGINEERING

Improving Airport Flight Prediction System Based on Optimized Regression Vector Machine Algorithm

Baraa Yousif Salman[1], Jaber Parchami[2*]

[1]Engineering Technical College, Imam Reza International University, Mashhad, Islamic Republic of Iran

[2]Department of Electrical Engineering, Sadjad University of Technology, Mashhad, Islamic Republic of Iran

[*] Corresponding author E-mail: jaber.parchami@sadjad.ac.ir

Article Info.

Article history:

Received
12 March 2024

Accepted
09 June 2024

Publishing
30 September 2024

Abstract

Nowadays, the safest and most extensive type of transportation is air flights, and tens of thousands of flights are carried out from different airports every day around the world. However, many of these flights have short-term or long-term delays, which cause the passengers' plans to change, and as a result, the airlines suffer losses. the first step to deal with these air delays is to predict them air delay. Regression functions and machine learning algorithms have a very good performance in predicting time series data. In this research, the optimized support vector regression (SVR) algorithm has been used to improve the accuracy of air delay prediction. The SVR algorithm is a machine learning algorithm that uses regression functions. The SVR algorithm has hyperparameters that play a key role in its performance and setting these parameters is very important. In this research, to improve the performance of the SVR algorithm in predicting air delays, the particle swarm optimization (PSO) algorithm has been used to adjust the hyperparameters of the SVR algorithm. The database used in this research was taken from the US Air Transport Fleet website and is an unbalanced database. Based on the results obtained from the simulations, our proposed method has improved with an average accuracy of 87.41% compared to other compared works.

Publisher: Middle Technical University

Keywords: Flight Delay; Prediction; Machine Learning; SVR Algorithm.

1. Introduction

Delay is a significant performance metric in transportation systems that individuals tend to remember clearly. In commercial aviation, delay is defined as the period during which an aircraft is kept from proceeding. A discrepancy between the planned and real departure or arrival times of an aircraft may suggest a delay [1]. Various statistics related to tolerance standards for flight delays are accessible to national regulatory authorities. Flight delay is a crucial topic in aviation transportation networks. In 2013, 16.3% of flights in Brazil had either cancellations or delays over 30 minutes. 36% of flights in Europe experienced delays over five minutes, compared to 31.1% of aircraft in the US facing delays exceeding fifteen minutes [2, 3]. This demonstrates the significance of the indicator and its impact regardless of the size of the meshes.

Airline delays have detrimental effects on travelers, airlines, and airports, mostly from an economic standpoint. Due to the unpredictability of their occurrence, travelers typically budget for longer journey times to get to their meetings on time [4, 5]. However, airlines face fines, penalties, and increased operating costs, including the cost of keeping crew and aircraft at airports [6, 7]. Delays can have negative environmental impacts by leading to higher fuel use and increased petrol emissions, as viewed from a sustainability standpoint [8–10].

Airlines rely on consumer loyalty for their frequent flyer programs, and reliable performance impacts consumer choice. Delays can jeopardize airlines' marketing efforts. a correlation has been seen between the frequency of flights, aircraft sizes, fares, and degrees of delays and complaints regarding airline service [11–13]. Estimating flight delays can help airport and airline officials make better tactical and operational decisions. It can also alert travelers so they can change their plans [14].

Large volumes of data from commercial aviation are collected and stored in databases every second to gain a comprehensive insight into the entire flight system. Engrossed in the copious amounts of data produced by sensors and the Internet of Things [15–17], Analysts and data scientists are improving their computational and data management abilities to derive value from every data point. Understanding the field, organizing data, and implementing a model are key components of "Data Science." Addressing problematic Big Data-related challenges is a typical approach.

This research has tried to provide an efficient prediction model to increase the accuracy of predicting the delay in the flight of airplanes. Using the results of this model in the future, researchers can find solutions to eliminate these delays as much as possible.

Nomenclature & Symbols			
PSO	Particle Swarm Optimization	GRA	Gray Relational Analysis
SVR	Support Vector Regression	GA	Genetic Algorithm
US	United States	BP	Backpropagation
DOT	Department of Transportation	LM	Lunberg-Marquardt
BTS	Bureau of Transportation Statistics	DL	Deep Learning
RUS	Random Under-Sampling	SAE	Stack Noise Removal Auto-Encoder
SVM	Support Vector Machine	SDA	Denoising Autoencoder
RNN	Recurrent Neural Network	BC	Betweenness Centrality
CAM	Cluster Airport Modeling		

This paper suggests a novel data-mining and machine-learning method for predicting delays in flight planning more accurately and preventing delays from escalating. This method utilizes the Particle Swarm Optimization (PSO) algorithm to enhance Support Vector Regression (SVR). The methodology is tested using the US flight networks.

Our motivation for using the PSO-SVR model and improving the SVR algorithm by the PSO algorithm to predict aircraft flight delays is to have a simple model with low computational complexity and optimal accuracy. In many of the models proposed in the past, very complex deep learning networks have been used for this purpose, which are not cost-effective either in terms of time or computation.

In the following, a summary of the literature is given in Section 2, and a suggested strategy for predicting aircraft delays is explained in Section 3. The experimental datasets and findings are described in Section 4. Lastly, the findings are given in Section 5.

2. Related Work

The authors in reference [18] Analyse network data to estimate aircraft delays at an airport and compare delay patterns with similar airports in the network. The "Cluster Airport Modelling" (CAM) technique uses network-based features as regressors to generate a time series for each cluster of airports and applies a common model such as a regression model with ARIMA autoregressive integrated moving average () (REG-ARIMA) to each cluster. Each model is applied separately to the data from each airport to forecast flight delays. a network analysis of airports was conducted and found that the betweenness centrality (BC) score was a useful predictor for anticipating flight delays. An investigation of 305 US airports and seven years of flight data demonstrates that CAM can accurately forecast flight delays.

In [19] is proposed a CNN-LSTM deep learning framework to consider the spatial–temporal correlations together with the extrinsic features for flight delay prediction. The CNN-LSTM model consists of a Convolution neural network (CNN) architecture to learn the spatial correlations followed by a Long short-term memory (LSTM) architecture to capture the temporal correlations. The spatial–temporal correlations obtained from the CNN-LSTM framework are then fused with the extrinsic features (e.g., airline issues, distance, schedule fly time, etc.) as inputs of the random forest (RF) model for flight delay prediction.

To estimate the likelihood of flight delays in aviation using aggregate data, the article [20] performs data mining employing causal machine learning algorithms in the USELEI (Understanding, Sampling, Exploring, Learning, Evaluating, and Inferring) process (perception, sampling, exploration, learning, evaluation, and inference). gathered from many sources. The results showed that the chance of a flight delay was significantly influenced by predictors such as reported arrivals and departures, demand for arrivals and departures, capacity, efficiency, and traffic volume at origin and destination airports. More significantly, how these predictors interact with one another and cause delayed events offers the causal interrelationships between variables in a properly structured network. Finally, to assess many what-if situations and create practical plans to lower the risk of delay, sensitivity analysis, and causal inference might be used.

The suggested model is examined in the paper [21] by combining various sampling strategies with machine learning methods. The model's practical application was illustrated using highly noisy, imbalanced, scattered, and skewed historical data from an international airline that operated in Hong Kong. According to the results, a machine learning algorithm using Constructive Neural Networks with an Artificial Minority Method Sampling Technique on Tomek's Sampling-Links Method (SMOTETomek) can classify the delay situation and predict the delay duration at the threshold of 60 minutes and 30 minutes, respectively, with a more balanced average recall accuracy of 65.5%, 61.5%, and 59% for a 4-hour forecast horizon. The suggested model outperformed the parallel model in terms of precision recall and area under the curve for minority labels, with scores of 32.44% and 35.14% compared to 26.43% and 21.02% for 60-minute and 30-minute thresholds, respectively. Additionally examined is the effect of various sampling methodologies, strategies, and estimating processes on forecast performance.

A deep learning (DL)-based model for flight delay prediction is presented in the article [22]. One of the newest techniques for handling complex problems with lots of data is deep learning (DL). Furthermore, significant features can be automatically extracted from the data by DL. Furthermore, a strategy based on the stack noise removal auto-encoder is devised and implemented in the proposed model, taking into account that noise is present in the majority of the flight delay data. The Lunberg-Marquardt algorithm is also utilized to determine the proper weight and bias values, and the output is then optimized to yield highly accurate results. Two additional structures have also been built to examine the impact of the LM algorithm and the stack noise elimination automatic encoder on the model's structure. The autoencoder and LM algorithm (SAE-LM) form the basis of the first structure, whereas the denoising autoencoder (SDA) alone forms the basis of the second. The model is applied to a US flight data set to analyze those three theories that may account for the imbalance in the data. Data correction was done using the small sample method in a balanced data set. These three models have been used for such tasks and their performance in terms of F-measure, memory, sensitivity, accuracy, and precision has been investigated in two different scenarios. The proposed prediction model's accuracy was compared to that of the previous prediction technique in terms of accuracy and tested the performance of the new model. The SDA-LM model becomes the most accurate when it is executed on proportionally balanced and unbalanced datasets. It surpasses both SAE-LM and the SDA models with accuracy, precision, sensitivity, recall, and F measure. The results indicate that the proposed model exceeds the performance of the previous RNN model in terms of accuracy evaluated on the different datasets. The Unger dataset is also performing better.

In [23] is proposed a framework integrated network called 'Attention-based Bidirectional long short-term memory' (ATT-BI-LSTM) for flight delay prediction. The Bidirectional LSTM model extracts the spatial and temporal of the flight network with weather features. The 'Attention mechanism' has been proposed to enable the model to discover significant and discriminating features that contribute to categorization. The first stage of the proposed framework is the 'preprocessing of dataset' which is performed through two steps. The first step is data transformation using a MinMax scaler to reduce the variation in the data. The second step is 'balancing the dataset' using the SMOTE technique for balancing data.

In [24], the study of the problem of flight delay propagation and prediction is modified. Firstly, the CNN-based CondenseNet algorithm is used to predict the delay level of the three-level flight chain data. Based on this, the CondenseNet network is improved by inserting CBAM modules and named CBAM-CondenseNet. The experimental results show that the improved algorithm can effectively improve the network performance, and the prediction accuracy can reach 89.8%. Compared with the traditional machine learning method, the average prediction accuracy increased by 8.7 percentage points.

Using supervised machine learning algorithms, the prediction model for airline arrival delays is shown in [25]. The forecast model was trained using meteorological data from July 2019 to December 2019 and domestic flight data from the United States. Flight delay prediction is the goal of the prediction model that was created using XGBoost and linear regression methods. Every algorithm's performance was examined. Weather and flight data were fed into the model. The XGBoost model utilised this dataset for binary classification to predict the presence of an arrival delay. A linear regression model was employed to forecast the length of the flight delay.

The study [26] introduces a new flight delay prediction model and explores the variables that could impact flight delay. By using the K-means clustering algorithm in the aircraft delay database, five alarm levels are defined. Using the Gray Relational Analysis (GRA) technique to identify important elements relevant to flight operations, an enhanced machine learning system known as GRA-Genetic technique (GA)-Backpropagation Neural Network Introduced is GRA-GA-BP, an optimized Is. by GA. The computation results demonstrate that the suggested prediction model based on GRA-GA-BP algorithm exhibits stronger stability and higher prediction accuracy when compared to the pre-optimization models and two other methods in earlier works. It also operates well in terms of memory usage and performance. The analysis in this paper demonstrates that this model can assist airlines in taking proactive measures on flights exhibiting anomalous trends and can also serve as an effective early warning system for flight delays.

Two probabilistic forecasting algorithms—random forest regression and mixed density networks—are used in the study [27] to predict aircraft delays at a European airport. The algorithms are effective and have an estimate of the desired probability distribution obtained with a mean absolute inaccuracy of under 15 minutes. multivariate probabilistic projections were introduced, including the expected delay, into the model which is designed to simulate a flight-to-gateway assignment problem to demonstrate the value of expected delay distribution. The task of waiting in the gate areas will be made more positive with this project. A comparison reveals that our aircraft-to-gate allocation model is more favorable compared to the deterministic aircraft-to-gate model with the number of aircraft involved in the project being decreased by 74% without considering possible delays. To sum up, it concludes adopting probabilistic forecasting within the provision of dependable airports operations is effective.

In [28], a new method is proposed to address these issues. In the proposed method, a group of potential indicators related to flight delay is introduced, and a combination of ANOVA and the Forward Sequential Feature Selection (FSFS) algorithm is used to determine the most influential indicators of flight delays. To overcome the challenges related to large flight data volumes, a clustering strategy based on the DBSCAN algorithm is employed.

In [29], a novel approach is developed which is an optimisation-driven deep learning model for predicting flight delays by extending a state-of-the-art method, DeepONet. the Box-Cox transformation is utilised for data conversion with a minimal error rate. Also, a deep residual network is employed for the feature fusion before training our model. Furthermore, this research uses flight on-time data for flight delay prediction.

In [30], it leveraged the power of artificial intelligence and machine learning techniques to build a framework for accurately predicting flight delays. To achieve this, is collected flight information from September 2017 to April 2023, along with weather data, and performed extensive feature engineering to extract informative features to train our model. A comparative analysis of various popular machine learning architectures with distinctive characteristics is conducted, aiming to determine their efficacy in achieving optimal accuracy on the newly proposed dataset.

In this section, some of the latest researches in the field of airplane flight delay prediction were reviewed. In most of these researches, either machine learning methods such as neural networks have been used to predict delay or statistical methods such as ARIMA method. However the combination of statistical methods, machine learning, and optimization algorithms is seen as a gap in most of the past works, and this research has tried to solve this gap.

3. Proposed Method

The ultimate purpose of the study is to modify the Regression Vector Machine (SVR) method to ensure that it produces accurate results. The performance of forecasting locations in the model depends largely on the kind of learning it has acquired. In this research, the SVR hyperparameters are determined with the PSO algorithm. PSO is a quick and effective method utilized in solving both non-linear and linear optimization problems.

Fig. 1 shows the diagram of the proposed method.

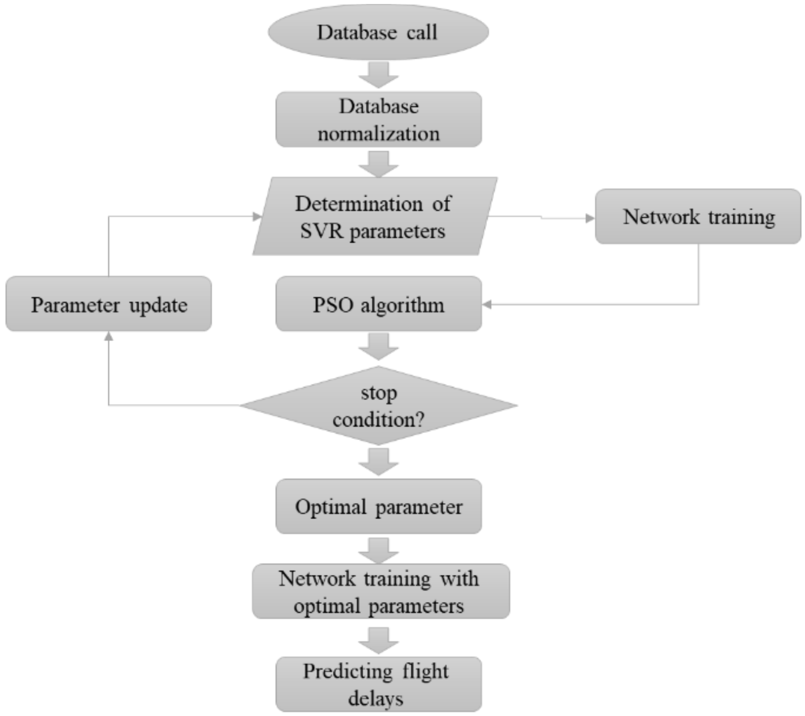

Fig. 1. Diagram of the proposed method

3.1. Data normalization

The following are considered in this step:

- Selecting a subset of the data
- Sort data for modeling
- Remove or replace missing values or empty values
- Providing a report on data quality

Before data mining, the database must undergo data cleansing and conversion to ensure it is in an acceptable format. Initially, the data is normalised to make it ready for input into the appropriate programme. Explanation of data normalisation will be provided at a later time.

Normalizing the data causes the data to be in an interval (L, H), for which the following relationship can be used

$$X_n = \frac{X_i - X_{min}}{X_{max} - X_{min}}(H - L) + L,\ i = 1,2, \dots, N \tag{1}$$

X_i represents the actual value of the network input, while X denotes the normalised value. X_min and X_max represent the least and maximum values of X, respectively. The lower limit (L) is -1 and the upper limit (H) is +1 in the normalisation interval. Therefore, the aforementioned relationship can be altered as follows.

$$X_n = \frac{2(X_i - X_{min})}{X_{max} - X_{min}} - 1 \tag{2}$$

The interval (L, H) is determined based on the problem type and the chosen transformation function, with a preference for using the intervals (0,1) or (+1, -1) over others. When the data is transformed to the range (0,1), the previous relation will be altered to the subsequent format.

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3}$$

3.2. SVR algorithm

Support Vector Regression (SVR) excels at ADDRESSING nonlinear issues and has been effectively applied across diverse domains [31]. Suppose the training sample is shown as $XY = \{(x.y)|(X_1.Y_1) \dots (X_{nd}.Y_{nd})\}$:

where *nd* is the number of training samples. In linear SVR, the relationship between input variable x_k and predictor variable $\hat{y}_k$ can be described by linear functions f(x) obtained from the following equation:

$$\hat{y}_k = f(x_k) = < w, x_k > + b \tag{4}$$

where • is the dot multiplication sign, w represents the weight vector, and b represents the bias vector. The objective is to identify a pair of unidentified vectors (w, b) that minimizes the forecast error for the training set and maximises the ε divergence from the true target y_k.

The second case means that there is no penalty for the pair of vectors during the optimization when $|y_k - f(x)| \le \varepsilon$ and is defined by the sensitive loss function, l_ε, which can be expressed as follows [32]:

$$l_\varepsilon = |y - f(x)|_\varepsilon = max\{0. |y - f(x)| - \varepsilon\} \tag{5}$$

To achieve optimal SRM and minimise complexity risk, it is important to reduce the value of soft w, where w^2=Thus, in mathematical words, the regression problem can be expressed as an optimisation problem in the following manner.

$$\min_{w,b,\xi_k,\xi_k^*} \frac{1}{2}|w^2| + C\sum_{k=1}^{nd} (\xi_k + \xi_k^*) \tag{6}$$

$$\text{subject to} \begin{cases} y_k - \langle w, x_k \rangle - b \leq \varepsilon + \xi_k \\ \langle w, x_k \rangle + b - y_k \leq \varepsilon + \xi_k^* \\ \xi_k, \xi_k^* \geq 0 \end{cases} \tag{7}$$

where ξ_k^* and ξ_k are auxiliary variables in equation 6, the constant parameter $C \geq 0$ Specifies the distance between the function's complexity and the deviation from the determined tolerable error ε in the previous stage. Equations (6) and (7) represent a quadratic optimisation issue that can be transformed into a Lagrange function. Maximise the dual optimisation problem to solve this Lagrange function.

The final solution to the optimization problem is:

$$\sum_{k=1}^{nd} (\alpha_k - \alpha_k^*)x_k \overset{yields}{\rightarrow} \hat{y}_{new} = f(x_{new}) = \sum_{k=1}^{nd} (\alpha_k - \alpha_k^*)\langle x_k, x_{new} \rangle + b \tag{8}$$

where $\alpha_k^* \geq 0$ and $\alpha_k \geq 0$ are Lagrange coefficientsThere is no text. Equation (8) shows that w can be fully defined as a linear combination of training vectors and Lagrange coefficients. The samples within the ε-insensitive region result in Lagrange coefficients being zero and are derived from training vectors outside this region. The problem's complexity is independent of its dimensions, as support vectors determine the function's complexity. To enhance the SVR algorithm for handling models with intricate non-linear relationships between input and output fields, pre-processing methods for training patterns can be employed. Mapping input vectors to a higher-dimensional space can be achieved by utilising kernel functions. Here is the solution:

$$\hat{y}_{new} = f(x_{new}) = \sum_{k=1}^{nd} (\alpha_k - \alpha_k^*)k\langle x_k, x_{new} \rangle + b \tag{9}$$

The kernel function maps the input space to a special space by weighting the points close to the data to create an estimate; Therefore, kernels are important to control the complexity of the final solution. In the strongly non-linear space, the RBF kernel usually shows more suitable results compared to the other mentioned kernels [33]. As a result, the RBF kernel is used in this article.

3.3. PSO algorithm

Particle swarm optimization is an optimization tool developed by Eberhart and Kennedy that was inspired by the behavior of bird swarms [34]. It is a crowd intelligence method that is commonly employed in numerical optimization issues and has acquired considerable popularity in recent years due to solving efficiency and effectiveness difficulties in the field of science and engineering. It is comparable to the PSO genetic algorithm based on a randomly initialized population. PSO searches from the ideal generation to the updated generation. Each component in PSO moves at a set speed. The velocity vector delivers motion to a particle whose motion is updated by two behavioral variables, namely memory (cognitive behavior) and present knowledge (social behavior) of each unique particle (bird). It can be predicted that enough time (repetition) of particles/birds will gather in the best spots they need. The foregoing behaviors, which are crucial to PSO, are expressed as follows:

$$v_i(j+1) = w(j)v_i(j) + \varphi_1(j)\left(pbest_i(j) - x_i(j)\right) + \varphi_2(j)\left(gbest_i(j) - x_i(j)\right) \tag{10}$$

$$\varphi_1(j) = C_1r_1(j), \varphi_2(j) = C_2r_2(j) \tag{11}$$

$$x_i(j+1) = x_i(j) + v_i(j+1) \tag{12}$$

where i represents the index of particles, j represents the repetition index, x_i represents the position of the particle, v_i represents the speed of the corresponding particles, pbest is the previous best position of the particle. gbest is the previous best position of the entire swarm. φ is a parameter that controls the speed value and is called contraction coefficient. In equation 10, the second term on the right side is a cognitive relationship and the third term is a social relationship. $C_1$and $C_2$are cognitive and social constants, respectively. r_1 and $r_2$are two random variables in the range (0.1). In some sources, *gbest* is specified individually for each particle with a defined proximity. In terms of neighborhood, each particle includes several other particles that affect the movement of particles. Neighborhood can be defined in different ways. Different criteria create different topologies and directly affect the results [35, 36].

3.4. PSO-SVR algorithm for flight prediction

Optimization in SVR algorithm is different from methods like neural network or regression. In SVR, the kernel function is a hyperparameter that determines the degree of similarity of various features, and as a result, they weight the optimizer functions. Another hyperparameter in SVR is the tuning parameter C, which is responsible for controlling the exchange between hyperplanes. The c adjustment parameter minimizes the training error. In the SVR algorithm, hyperparameters play the main role in the accuracy of the algorithm, and normally these hyperparameters are assigned randomly.

Meta-heuristic algorithms can be effective in the optimal setting of SVR hyperparameters. One of the most famous and efficient meta-heuristic algorithms is the PSO algorithm, whose mathematical and theoretical explanations were examined in section 3.3. In this work, the PSO algorithm is used to optimize SVR hyperparameters, the structure of which is shown in Fig. 2. PSO minimizes the prediction error in the SVR algorithm and increases the accuracy. The working method is that in each repetition of the algorithm, the prediction error of the model is obtained, and then the parameters are adjusted again, and in the next repetition, the error is obtained again, and in this way, in several steps, the hyperparameters that obtain the least error are selected, and then the network with optimal hyperparameters is tested on test data and flight delay predictions are made.

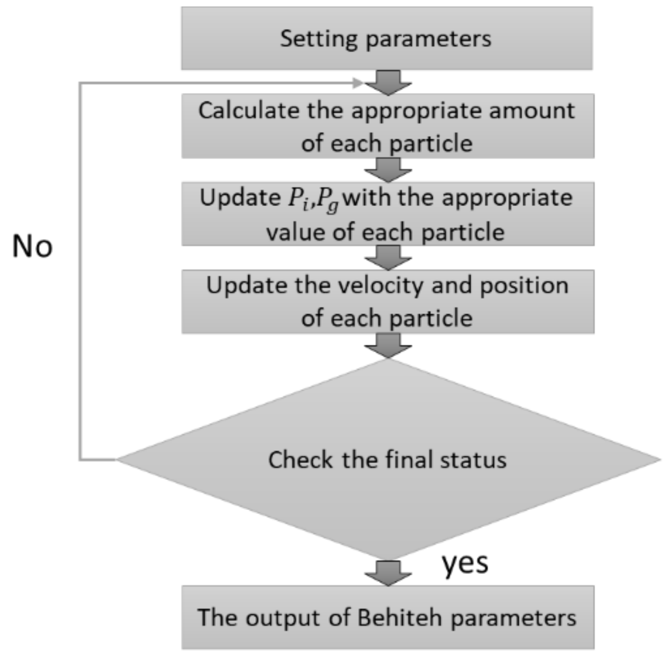

Fig. 2. Optimization structure of SVM parameters using PSO

4. Evaluation of Simulation Results

Our main goal in this work is to predict air delays by airlines using the SVR algorithm optimized by the PSO algorithm, as described in section 3. Whether a delay occurs or not is the subject of this prediction. The results of the implementation of the proposed method in MATLAB version 2021 are presented in this section. In the rest of this section, numerical explanations for each component of the proposed model are presented, numerical results of the proposed technique are presented, and then a comparison between the proposed method and some other methods mentioned in the reference article is presented.

4.1. Database

The Bureau of Transportation Statistics (BTS), a division of the US Department of Transportation (DOT), monitors the punctuality of domestic flights conducted by major airlines. The Department of Transportation's monthly Air Travel Consumer Report details the number of flights that are on time, delayed, cancelled, and redirected. The report is usually published about 30 days after the end of the month and can be viewed in a summarized version on this page. The site is visible.

The information needed in this thesis was taken from this site. Also, all the information and this database itself is available on the Kaggle website [37].

4.2. Evaluation criteria

In this work, four criteria of accuracy, correctness, readability and F1 score are used. Using these four criteria, the proposed method has been evaluated. Below are the relationships of these four criteria:

$$Accuracy = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ Population} \tag{13}$$

$$Precision = \frac{\sum True\ Positive}{\sum Test\ Outcome\ Positive} \tag{14}$$

$$Recall = \frac{\sum True\ Positive}{\sum Realy\ Positive} \tag{15}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{16}$$

4.3. Unbalanced data sampling

The nature of the databases this paper examines is imbalanced, and latency records are uncommon. A data collection is considered "imbalanced" if it contains a rare class that is much more important than the other classes. The class with more samples in an unbalanced data set is referred to as the main class, while the class with comparatively fewer samples is referred to as the subclass. We are compelled to employ a sample strategy to overcome this issue because delayed flights are included in the subclass in our database. Sampling is a useful strategy for handling uneven data. Because conventional classification algorithms do not account for class imbalance in their operations, they perform poorly when a dataset is uneven. Regardless of whether the data are from minority or template classes, classical techniques handle them in the same manner. Maximizing overall accuracy is their aim. The generalization of machine learning outcomes is decreased when consideration is given to the relative distribution of every class in unbalanced data. In this work, specifically is employed random subsample sampling (RUS). This technique involves copying and dispersing uncommon samples across a data collection while randomly removing examples of a majority class from the data set.

4.4. SVR parameter setting

A typical SVR algorithm selects SVR meta-parameters at random. To determine the ideal SVR parameter values, PSO, a general optimization technique, is employed in this situation. PSO looks for an optimal point across the whole response space, as previously mentioned. Every point in the response space has the potential to be the ideal value. PSO has the benefit of reducing the time needed to find optimal values by utilizing a clever algorithm that eliminates the need to look at every point in the solution space

4.5. Results of the proposed method

The proposed algorithm should be evaluated on a different data set to obtain an accurate and appropriate overview of the proposed method. 30% of the data was considered for testing and evaluating the proposed method, and 70% was allocated for training the proposed algorithm and locating the optimal values in SVR. Simulation experiments have been performed using the proposed PSO-SVR method. At first, the proposed model completes the learning by using the data set of the training section. And they are validated using the test data set to check the accuracy of the prediction models used. The main criterion in this work is prediction accuracy.

As explained above, the main step in this method is repeating the PSO algorithm step by step to find the best hyperparameters in the SVR algorithm. Fig. 3 shows the convergence diagram of the PSO algorithm. Based on the convergence diagram, the PSO algorithm has converged in 5 steps

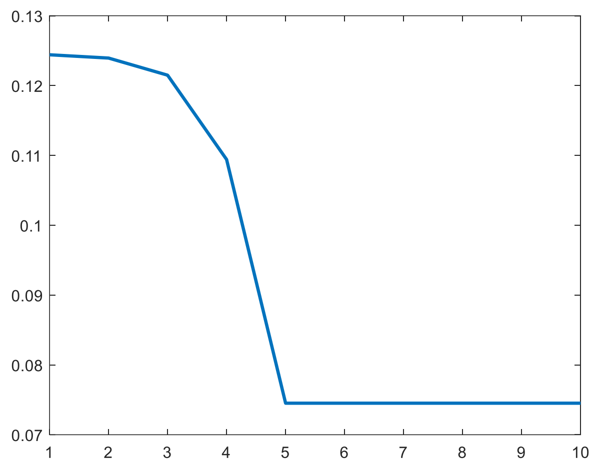

Fig. 3. PSO algorithm convergence diagram

Next, in Fig. 4 the confusion matrix for the accuracy of predicting the occurrence of delay is presented. According to this figure, the overall accuracy of the proposed method is 91.9%, which is very accurate for this type of data. It should be noted that this accuracy is for a single execution of the program, and to achieve a valid accuracy, an average of dozens of executions must be taken.

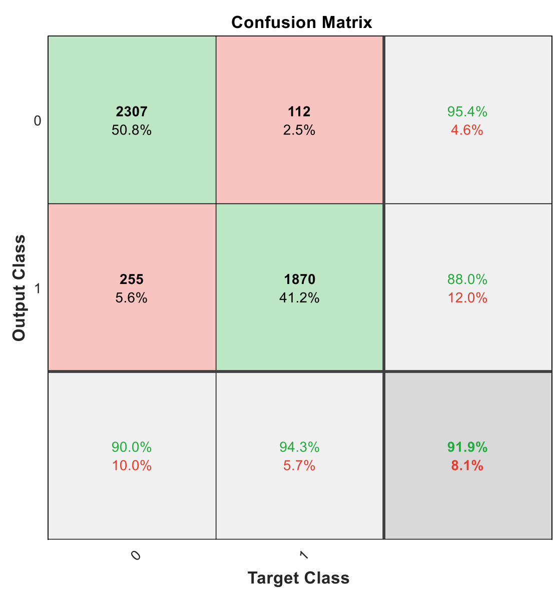

Fig. 4. Confusion matrix for the proposed method

In Fig. 5, the graph of other evaluation criteria such as recall, accuracy and F-score is calculated and displayed.

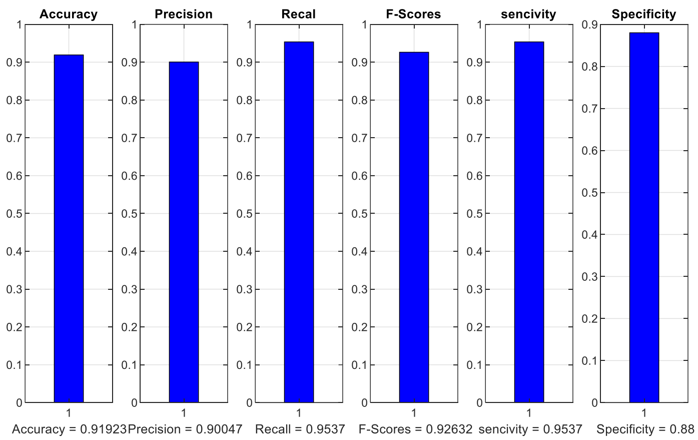

Fig. 5. Different criteria for the proposed method

4.6. Comparing results

In this part, the proposed method is compared with other methods presented in the reference article in terms of accuracy. Since in unbalanced data, the samples in some classes are very few, as a result, machine learning models have problems in the training stage and the average prediction accuracy is greatly reduced. Table 1 shows the comparison between other methods. The level of accuracy mentioned in this table represents the accuracy for the average of 50 program execution times.

Table 1. Comparing the results of the reference article with the proposed method

Method	Accuracy (%)
Decision tree [38]	70.87
Cluster classification [38]	69.63
Hybrid method [38]	76.44
Bayesian approach [38]	70.17
Random forest [38]	72.11
CNN-LSTM [19]	84
Proposed Method	87.41

As it is known, the proposed method has much better accuracy than other methods. This precision is very appropriate considering the unbalanced nature of the data.

5. Conclusion

One of the most essential steps in appropriate and structured planning for passenger satisfaction in air transportation networks is the prediction of flight delays. It is important to determine the kinds and quantities of delays to optimize the air transportation network's resilience to interruptions. In this paper, the flight delay is predicted by the PSO-SVR algorithm, and the results were compared with other machine learning methods. In the proposed model of this research, the PSO algorithm is used to optimize the hyperparameters of the SVR algorithm. This optimization improves the performance of the SVR algorithm in predicting the delay of airplane flights. The obtained results showed the appropriate accuracy of the proposed method in delay prediction. Parameters such as fleet age and aircraft type as well as weather conditions have strong effects on flight delays in the airline network. These results can be useful for planning in the air transport fleet to reduce the amount of delay.

Acknowledgment

I would like to thank the Engineering Technical College, Imam Reza International University, for supporting the project requirements.

References

[1] F. Wieland. Limits to growth: results from the detailed policy assessment tool [air traffic congestion]. In 16th DASC. AIAA/IEEE Digital Avionics Systems Conference. Reflections to the Future. Proceedings, volume 2, pages 9.2–1–9.2–8 vol.2, Oct. 1997. DOI: 10.1109/DASC.1997.637296.

[2] Eurocontrol. Coda Digest - Delays to Air Transport in Europe. Technical report, https://www.eurocontrol.int/articles/coda-publications, 2017.

[3] ANAC. Agencia Nacional de Aviac¸ ˆ ao Civil. Technical report, http: ˜ //www.anac.gov.br/, 2017.

[4] P. Balakrishna, R. Ganesan, and L. Sherry. Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures. Transportation Research Part C: Emerging Technologies, 18(6):950–962, Dec. 2010. ISSN 0968-090X. https://doi.org/10.1016/j.trc.2010.03.003.

[5] P. Fleurquin, B. Campanelli, V. Eguiluz, and J. Ramasco. Trees of reactionary delay: Addressing the dynamical robustness of the US air transportation network. In SIDs 2014 - Proceedings of the SESAR Innovation Days, 2014.

[6] R. Britto, M. Dresner, and A. Voltes. The impact of flight delays on passenger demand and societal welfare. Transportation Research Part E: Logistics and Transportation Review, 48(2):460–469, Mar. 2012. ISSN 1366- 5545. https://doi.org/10.1016/j.tre.2011.10.009.
[7] Y. Tu, M. O. Ball, and W. S. Jank. Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. Journal of the American Statistical Association, 103(481):112–125, 2008. https://doi.org/10.1198/016214507000000257.
[8] T. Krstic Simi ´ c and O. Babi ´ c. Airport tra ´ ffic complexity and environment efficiency metrics for evaluation of ATM measures. Journal of Air Transport Management, 42(Supplement C):260–271, Jan. 2015. ISSN 0969- 6997. https://doi.org/10.1016/j.jairtraman.2014.11.008.
[9] E. Balaban, I. Roychoudhury, L. Spirkovska, S. Sankararaman, C. Kulkarni, and T. Arnon. Dynamic routing of aircraft in the presence of adverse weather using a POMDP framework. In 17th AIAA Aviation Technology, Integration, and Operations Conference, 2017, 2017. https://doi.org/10.2514/6.2017-3429.
[10] Y. Xu, R. Dalmau, and X. Prats. Maximizing airborne delay at no extra fuel cost using linear holding. Transportation Research Part C: Emerging Technologies, 81:137–152, 2017. https://doi.org/10.1016/j.trc.2017.05.012.
[11] W. Cong, M. Hu, B. Dong, Y. Wang, and C. Feng. Empirical analysis of airport network and critical airports. Chinese Journal of Aeronautics, 29(2):512–519, 2016. https://doi.org/10.1016/j.cja.2016.01.010.
[12] V. Pai. On the factors that affect airline flight frequency and aircraft size. Journal of Air Transport Management, 16(4):169–177, July 2010. ISSN 0969-6997. https://doi.org/10.1016/j.jairtraman.2009.08.001.
[13] B. Zou and M. Hansen. Flight delay impact on airfare and flight frequency: A comprehensive assessment. Transportation Research Part E: Logistics and Transportation Review, 69(0):54 – 74, 2014. ISSN 1366-5545. https://doi.org/10.1016/j.tre.2014.05.016.
[14] A. D'Ariano, M. Pistelli, and D. Pacciarelli. Aircraft retiming and rerouting in the vicinity of airports. IET Intelligent Transport Systems, 6(4):433–443, 2012. DOI: 10.1049/iet-its.2011.0182.
[15] M. Mellat-Parast, D. Golmohammadi, K. McFadden, and J. Miller. Linking business strategy to service failures and financial performance: Empirical evidence from the U.S. domestic airline industry. Journal of Operations Management, 38:14–24, 2015. https://doi.org/10.1016/j.jom.2015.06.003.
[16] B. Campanelli, P. Fleurquin, V. Egu´ıluz, J. Ramasco, A. Arranz, I. Extebarria, and C. Ciruelos. Modeling reactionary delays in the European air transport network. In SIDs 2014 - Proceedings of the SESAR Innovation Days, 2014.
[17] E. R. Mueller and G. B. Chatterji. Analysis of aircraft arrival and departure delay characteristics. In AIAA aircraft technology, integration and operations (ATIO) conference, 2002. DOI:10.2514/6.2002-5866.
[18] Güvercin M, Ferhatosmanoglu N, Gedik B. Forecasting flight delays using clustered models based on airport networks. IEEE Transactions on Intelligent Transportation Systems. 2020 May 11;22(5):3179-89. DOI: 10.1109/TITS.2020.2990960.
[19] Li Q, Guan X, Liu J. A CNN-LSTM framework for flight delay prediction. Expert Systems with Applications. 2023 Oct 1;227:120287.
[20] Truong D. Using causal machine learning for predicting the risk of flight delays in air transportation. Journal of Air Transport Management. 2021 Mar 1;91:101993. Li, Qiang and Guan, Xinjia and Liu, Jinpeng, A CNN-LSTM Framework for Flight Delay Prediction (November 22, 2022). Available at SSRN: https://ssrn.com/abstract=4283680 or http://dx.doi.org/10.2139/ssrn.4283680.
[21] Khan WA, Ma HL, Chung SH, Wen X. Hierarchical integrated machine learning model for predicting flight departure delays and duration in series. Transportation Research Part C: Emerging Technologies. 2021 Aug 1;129:103225. https://doi.org/10.1016/j.trc.2021.103225.
[22] Yazdi MF, Kamel SR, Chabok SJ, Kheirabadi M. Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. Journal of Big Data. 2020 Dec;7:1-28. https://doi.org/10.1186/s40537-020-00380-z.
[23] Mamdouh M, Ezzat M, Hefny H. Improving flight delays prediction by developing attention-based bidirectional LSTM network. Expert Systems with Applications. 2024 Mar 15;238:121747. https://doi.org/10.1016/j.eswa.2023.121747.
[24] Qu J, Wu S, Zhang J. Flight Delay Propagation Prediction Based on Deep Learning. Mathematics. 2023; 11(3):494. https://doi.org/10.3390/math11030494.
[25] N. L. Kalyani, G. Jeshmitha, B. S. Sai U., M. Samanvitha, J. Mahesh and B. V. Kiranmayee, "Machine Learning Model - based Prediction of Flight Delay," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2020, pp. 577-581, doi: 10.1109/I-SMAC49090.2020.9243339.
[26] Shi T, Lai J, Gu R, Wei Z. An improved artificial neural network model for flights delay prediction. International Journal of Pattern Recognition and Artificial Intelligence. 2021 Jun 30;35(08):2159027. https://doi.org/10.1142/S0218001421590278.
[27] Zoutendijk, Micha, and Mihaela Mitici. 2021. "Probabilistic Flight Delay Predictions Using Machine Learning and Applications to the Flight-to-Gate Assignment Problem" Aerospace 8, no. 6: 152. https://doi.org/10.3390/aerospace8060152.
[28] Dai M. A hybrid machine learning-based model for predicting flight delay through aviation big data. Sci Rep. 2024 Feb 26;14(1):4603. doi: 10.1038/s41598-024-55217-z. PMID: 38409455; PMCID: PMC10897135.
[29] Bisandu DB, Moulitsas I. Prediction of flight delay using deep operator network with gradient-mayfly optimisation algorithm. Expert Systems With Applications. 2024 Aug 1;247:123306. https://doi.org/10.1016/j.eswa.2024.123306.
[30] Alfarhood M, Alotaibi R, Abdulrahim B, Einieh A, Almousa M, Alkhanifer A. Predicting Flight Delays with Machine Learning: A Case Study from Saudi Arabian Airlines. International Journal of Aerospace Engineering. 2024;2024(1):3385463. DOI: 10.1155/2024/3385463.
[31] Ghasemi, E., Kalhori, H. & Bagherpour, R. A new hybrid ANFIS–PSO model for prediction of peak particle velocity due to bench blasting. Engineering with Computers 32, 607–614 (2016). https://doi.org/10.1007/s00366-016-0438-1.
[32] Vapnik, V. N. and Vapnik, V., Statistical learning theory Vol. 1: Wiley New York, 1998.
[33] Gunn, S. R., "Support vector machines for classification and regression," ISIS technical report, Vol. 14, 1998.
[34] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, 1995, pp. 39-43, doi: 10.1109/MHS.1995.494215.
[35] J. Kennedy, "Bare bones particle swarms," Proceedings of the 2003 IEEE Swarm Intelligence Symposium. SIS'03 (Cat. No.03EX706), Indianapolis, IN, USA, 2003, pp. 80-87, doi: 10.1109/SIS.2003.1202251.
[36] Mendes, R., Kennedy, J., and Neves, J., "The fully informed particle swarm: simpler, maybe better," IEEE transactions on evolutionary computation, Vol. 8, pp. 204-210, 2004. DOI:10.1109/TEVC.2004.826074.
[37] Airlines Delay, Airline on-time statistics and delay causes, https://www.kaggle.com/datasets/giovamata/airlinedelaycauses.
[38] H. Khaksar, A. Sheikholeslami, "Airline delay prediction by machine learning algorithms", Sharif University of Technology, Scientia Iranica, Transactions A: Civil Engineering, (2019) 26(5), 2689- 2702. Doi: 10.24200/SCI.2017.20020.