# Literature Survey

To predict flight delays using machine learning, you will need to collect and process a large amount of data on past flight delays. This data should include information such as the flight's departure and arrival times, the airline, the aircraft type, and the weather conditions at the departure and arrival airports. Once you have collected and cleaned the data, you can use a variety of machine learning techniques such as regression, decision trees, or neural networks to train a model that can predict flight delays based on this data. It is important to note that flight delay prediction is a highly complex task and requires a lot of data.

The literature suggests that ML models, specifically decision tree, ANN and random forest models, have been used to predict flight delays with varying degrees of accuracy. Commonly used features include historical flight data, weather conditions, and airport operations. It also shows that a combination of data mining techniques can be used to identify the factors that contribute to flight delays.

In one of the best studies [56] that has been performed based on statistics delay time has been considered to be reduced. Their study has investigated important factors before fly and those which occur on the ground. In the next step, it has predicted the delay at destination based on factors that occur in the vicinity of arrival time at destination. Eventually, results have shown that whenever, the delay is correctly predicted, passenger disaffection and fuel consumption decrease and consequently number of flight increases. Moreover, it is possible to increase the agencies' benefits through reducing number of passengers who wrongly selected their routs or specifying the probabilities for some flights and optimizing delay time prediction.

Another prominent investigation based on Probability [57] has been done and the author believes that huge storm in U.S.A has highly affected the flight delay. This study has been devoted to predict delay based on mathematical calculations and through considering delay time duration of the flights that had been engaged to storm in the same day. Metrological reports have shown the effect of storm one hour before and after event cause ephemeral climate at the region. In the next step, Monte-Carlo simulation has been used to estimate the airport runway capacity, so that traffic of each runway would have been estimated. As the research has employed only one factor, the model has not enough accuracy, but it is possible to increase region air capacity path structure [57].

A model has been presented in [82], which is one of the best network-based models. The researchers have presented a model based on Bayesian and Gaussian mixture model- expectation maximization (GMM-EM) algorithm to predict and analyze the factors affecting the flight delay in Brazil for several point along the path. At the first stage of model, the degree of effectiveness for each factor is specified and then it has specified investigated whether the delay had happened

in a greater domain or no. the next delay probability is computed using GMM-EM [82] and EM algorithm which are specified based on similarity. The result has shown that it is possible to predict the probability of delay in higher levels through specifying low level factors. Moreover, GMM-EM [82] similarity function has more values rather than EM algorithm [82] in each step, so that the results would have been converged sooner. In addition, the model accuracy is increased, so that the prediction is more trustable.

One of the best studies [93] in the area of operating method has been presented. Studied the effects of capacity and damage on different levels of delay in American airports.

Other simulations focus on stability and reliability during the delay and its propagation. For instance, in [90] the problems of congestion were studied. Then, a queue-based model was presented for analyzing delay propagation in consecutive flights in the Los Angeles airport.

One of the best studies [119] in the area of machine learning method has been presented by a model which applicate machine learning techniques to investigate delay in arrival flights. This research firstly has extracted important characteristics and then has been used for both neural networks and deep believe network through arbitrary samples to train the model. The model utilizes Memento [119] and Resilient Back Optimized Propagation [119] that the Resilient back propagations quicker than back propagation [119] and as a result the model training and consequently has been increased. Deep believe networks [119] is based on a few Boltzmann machine [119] that each communication layer receives data from the previous layer and in each step a Boltzmann machine [119] is added to Believe Network overall, training time reduced using parameter adjustment operation and learning rate, false classified error rate. As each layer has convergence at the output, training speed is reduced and the gradient approaches zero. In addition, a relatively small data base is used for the model because of limited system capacity. So that this problem leads to a noticeable reduction in prediction precision whenever it is not at database.

A model has been presented [125] which was one of the machine learning method. the researcher has presented a model based on support vector regressor (SVR) algorithm to predict flight delay in U.S.A airports. Due to the large amount of data, the data was grouped and sampled by month. At the first stage for categorical variables, cat-boost used the ordered boosting method. Because cat-boost itself had the effect of scoring features, it was possible to select parameters that were more important to the model when the threshold was unknown, so cat-boost was used to evaluate the features of each feature to select features, and finally 15 features were selected to build a training model.

Then has been used several common regression prediction algorithms to predict the delay at the same time for the round-trip flight between John F. Kennedy International Airport and O'Hare International Airport.

Finally, the specific delay time was predicted. The results have shown SVR has the best prediction result for the flight delay time with the best accuracy value was 80.44%. Also, the time characteristics had a large impact on the mode performance.

The air time and flight distance would also have a greater impact on on-time performance of specific flight; Different carriers and specific aircraft would also have a slight influence of on time performance. Accuracy of this model is low because detailed weather and aircraft data could not be collected.

A research [126] analyzes flight information of U.S domestic flight operated by American Airlines, covering top 5 busiest airports of US and predicting possible arrival delay of the flight using Data Mining and Machine Learning Approaches. Due to the imbalanced data, Over-Sampling technique, Randomized SMOTE was applied for Data Balancing. The Gradient Boosting Classifier Model was deployed by training and then Grid Search on Gradient Boosting Classifier Model on flight data, caused hyper-parameter tuned and achieving a maximum accuracy of 85.73%. Result showed that deleting some features affected the value of accuracy and reduced it.

A group of researchers [127] have designed 5 models to predict flight delay based on machine learning models such as Logistic Regression, Decision Tree Regression, Bayesian Ridge, Regression and Gradient Boosting Regression. They collected data from Bureau of Transportation, U.S. Statistics of all the domestic flights taken in 2015 and predicted whether the arrival of a particular flight would be delayed or not.