LITERATURE SURVEY

With increasingly tight flight schedules, the prediction of aviation resources is developing rapidly. The differences in the current research are mainly in the prediction methods and the input factors considered. Prediction methods are either based on statistics (Stats) or based on machine learning (ML) or deep learning (DL). The influencing factors considered are mainly divided into direct and indirect factors. As mentioned earlier, the direct influencing factors are those that have nothing to do with the time series, which will not be accumulated. However, the indirect factors are related to the time series, these factors will accumulate over time, and finally affect the delay of a flight.

Much literature addresses the statistical analysis. Tu et al. used a genetic algorithm to fit delay data and study long- and short-term flight departure trends . The model included seasonal influences, daily trends, and random trends, enabling users to grasp general delay characteristics. Hsiao and Hansen considered the influence of arrival queues, passenger flow, weather and other factors on flight delays . Through econometric analysis of the contribution rates of various factors to delays, the model explained 72–73% of the variation in the average delay. Hao et al. used econometric and simulation models to calculate and decompose delays, considering direct factors such as quarter-hourly data on throughput, demand, and arrival rates. Rodriguez-Sanza et al. used a Bayesian network and time-series features to model randomness and time variation of flight delays. However, the prediction results consisted of statistical guidance rather than a tactical operation.

ML and DL are developing rapidly. Rebollo and Balakrishnan divided flight delay data into temporal data (e.g., day of week, month of year, and time of day) and spatial data (e.g., delay state and type of delay). They used a random forest model to predict departure delays in the next two hours, with an average error of about 21 min on the test dataset. However, they only conducted a sensitivity analysis of

influencing factors and not the primary factors (or time points) of delays. Manna et al. used the gradient boosted decision tree model (GBDT) to predict the delay of a flight with six directed factors. Their model can also be used by passengers or airline agencies. Kim et al. placed delays at different levels for prediction, using a recurrent neural network (RNN) to consider time and other direct factors such as weather and visibility. Based on a single airport, McCarthy et al. studied the delays of multiple airports with a long short-term memory (LSTM) algorithm, using the time series of the past 24 h to predict delays. The method was shown to be accurate and robust for low-cost airlines in Europe. The analysis of causes on the whole network is helpful to gain an overall understanding of critical factors. Qiang uses the Random Forest (RF) algorithm to predict the delay of a single airport , and the method was validated by U.S. domestic flights. However, it cannot explain the delay for one flight.

Zhen and Bin et al. combine the RF and the maximal information coefficient to analyse the flight delay of PEK, but there is no analysis of factors of delay in detail .

Ai and Pan et al. used a convolutional LSTM (Conv-LSTM) algorithm to analyse the flight delay distribution, considering indirect factors such as pre-order flight delay, route congestion and airport capacity. Because of the large amount of data and wide range of prediction time window, the study used relatively few factors. Hao et al. make a multi-step prediction of airport delay with spatio-temporal data , and the model used historical data of multiple airports for training. The prediction accuracy of the model is high, but it is unable to analyse a single flight.

Yu et al. combined a deep belief network and SVR to mine the influencing factors of a flight delay with better prediction results than benchmark methods such as k-nearest neighbours (KNN), support vector machine (SVM) and linear regression (LR).However, the paper used a statistical method to analyse critical factors and could not give a more intuitive, detailed, tactical description of each delay.

The analysis of flight delay and discovery of influencing factors are becoming more and more detailed. Liu et al. used machine learning methods including SVM, LR and RF to analyse the impact of convective weather, local weather and airport traffic demand on ground delay. Still, they could not guide specific work. Research on the prediction of delay is developing from two classifications to multi-classification. Gui and Liu combined several machine learning and deep learning methods with information on weather, flights, airports, and other direct influencing factors to classify and regress airport delays.

Jia and Honghai et al. combined KNN, RF, LG, Decision Tree and Gaussian Naïve Bayes with directed factors to predict flight delays of Boston Logan International Airport . The results showed that the stochastic forest model could achieve better prediction accuracy than baseline algorithms, but factors affecting prediction results were not discussed. Khan et al. integrates machine learning algorithms to study the delay and the duration of the delay, but they didn't consider indirect factors. Junfeng et al. built eight widely used modes including linear regression models, non-linear regression models, and tree-based ensemble models, and they found that if the feature set could capture the arrival characteristics, even simple linear regression models or algorithms could fulfil the prediction task.

With the improvement of computing power, the DL algorithm based on multi-airport data has received more and more attention . Multi-airport delay prediction needs to consider higher data dimensions, such as OD (origin-destination) data between airports, which will not be conducive to the updating of the model. At the same time, They are difficult to give a specific analysis for one specific flight.