# Prediction of Weather Impacted Airport Capacity using Ensemble Learning

*Yao Wang, NASA Ames Research Center, Moffett Field, California*

## Abstract

Ensemble learning with the Bagging Decision Tree (BDT) model was used to assess the impact of weather on airport capacities at selected high-demand airports in the United States. The ensemble bagging decision tree models were developed and validated using the Federal Aviation Administration (FAA) Aviation System Performance Metrics (ASPM) data and weather forecast at these airports. The study examines the performance of BDT, along with traditional single Support Vector Machines (SVM), for airport runway configuration selection and airport arrival rates (AAR) prediction during weather impacts. Testing of these models was accomplished using observed weather, weather forecast, and airport operation information at the chosen airports. The experimental results show that ensemble methods are more accurate than a single SVM classifier. The airport capacity ensemble method presented here can be used as a decision support model that supports air traffic flow management to meet the weather impacted airport capacity in order to reduce costs and increase safety.

## I. Introduction

The steady rise in demand for air transportation and the restricted capacities of the National Airspace System (NAS) increases the possibility of airspace congestion. At the major commercial airports, air traffic congestion has been a serious problem for Air Traffic Management (ATM) [1]. FAA Traffic Flow Management (TFM) manages air traffic flow to balance the air traffic arrival demand against airport arrival capacity during inclement weather or under other circumstances. This often results in reduced airport arrival capacity which causes airborne delays by holding some aircraft for landing or changing routes to stay in clear weather to maintain safety. At major airports in the United States, when the expected demand for arrival air traffic flow exceeds the airport capacity for a significant period of time, Ground Delay Program (GDP), as one of the Traffic Management Initiatives (TMI), will then be implemented to smooth out the arrival flow and bring arrival demand in line with airport capacity [2, 3].

GDP is the most commonly used air traffic management procedure where aircraft are delayed at their departure airport in order to balance demand and capacity at their arrival airport. During the GDP, flights are assigned new departure clearance times, and will receive delays that in turn control their arrival times at the impacted airport. These are very important because the airborne delay is being changed to ground delay, which is both less costly and less risky. The most common reason for an overage of demand versus capacity is the reduction in airport acceptance rate due to adverse airport weather, such as strong wind, low ceilings and low visibility.

For efficient GDP operation, accurate and reliable prediction of arrival demand and airport arrival capacity is crucial. TFM uses flight schedule monitor (FSM) software to compile scheduled flight information and flight plans in order to predict the demand for arrivals and departures at the site. During GDP, the major cause of the reduction in airport capacity is inclement weather. Since weather forecast products are often inaccurate and the uncertainty increases with forecast lead time, the problem of predicting airport arrival capacity, known as Airport Arrival Rate (AAR), is difficult to address.

AAR is a dynamic parameter specifying the number of arrival aircraft that can be landed at a given airport in a one-hour time frame [4]. The short-term forecast of capacity used by GDPs is usually given for each hour over several hours. For instance, AAR might be predicted over a six-hour period. ATM considers forecast weather conditions for an airport while making decisions about runway configurations and subsequent AARs. The main problem with prediction is due to highly stochastic nature of weather conditions that ultimately determine AARs. If the forecast AARs turn out to be higher than the AARs that actually materialized, then

unnecessary ground delay will be applied and valuable airport capacity will not be fully utilized. Similarly, if the forecast AARs are lower than the actual AARs, then demand will exceed capacity and there will be airborne holding that could have been replaced with ground holding which is more safe and costs less.

Besides weather, other factors, such as runway configuration, aircraft fleet mix, air traffic control separation requirements, arrival/departure split, as well as controller workload, etc. [5], also can affect airport runway arrival capacity. There are many runway capacity models available today [6]. Some models have been successful in improving ATM operations in some cases. However, due to their inability to accurately forecast airport weather impacts, a complementary weather decision support tool to translate weather forecast information into its impacts on airport capacity is important for TFM operation.

The airport capacity of multiple runway systems in a large airport is determined by the number of runways available for simultaneous use. Runway configuration (grouping of runways) is a critical factor in determining airport capacity. Runway configuration usage depends on airport weather conditions, noise abatement procedures, air traffic demand, airport operator constraints, surface congestion, and navigational system outages. Among these factors, the most significant one is weather, wind direction and speed in particular. Even though runway configuration selection is a critical element in air traffic flow management, current operations in runway management are without assistance from automation.

This paper examines the weather impacts on airport runway configuration selection and arrival capacity and introduces the use of Support Vector Machines (SVM) and Ensemble Bagging Decision Tree (BDT) machine learning method to select the runway configuration and predict the AARs for several hours. This approach relies on the historical airport operation and weather data to develop and test SVM and BDT models. A comparison of the model predictions is presented. The results from this approach are also discussed in this study.

The remainder of the paper is organized as follows. Section II describes SVM and BDT modeling machine learning approaches and discusses the methods used to model performance validation. Section III shows the experimental data setup. Then Section IV presents analysis and computational results on the estimation of runway configuration selection and prediction of AAR using SVM and BDT approaches for several major airports. Finally, concluding remarks are provided in Section V.

# II. Modeling Methodology

Support Vector Machine (SVM) and Bagging Decision Tree (BDT) were used in this study to estimate airport runway configuration usage and predict AAR using airport operation and weather information. The modeling approach represents a data-driven method for resolving classification tasks. Supervised machine learning was used to train BDT and SVM models by mapping inputs to desired outputs or targets. The models were validated using data cross validation methods.

### *SVM Classification*

The Support Vector Machine (SVM), a supervised machine learning algorithm, was invented by Vapnik et al. [7-9] and has been successively extended by a number of other researchers. Its robust performance with respect to limited, sparse and noisy data is making it widely used in many applications from protein function, and face recognition, to text categorization for classification and regression prediction. The SVM model has also been utilized in airport capacity classification prediction [10].

When used for binary classification, the SVM algorithm separates a given set of two-class training data by constructing a multidimensional hyper-plane that optimally discriminates between the two clusters. Although SVMs were originally proposed to solve linear classification problems, they can be applied to non-linear decision functions by using the so-called kernel function trick [11]. Adopting this kernel technique, SVM can be utilized to automatically realize a non-linear mapping to a high dimensional space. The hyper plane in the high dimensional space corresponds to a non-linear decision boundary in the input space. A widely used kernel is the Gaussian radial basis function (RBF). In this study, the SVM classification is implemented using LIBSVM [12].

### *Ensemble Bagging Decision Tree*

Ensemble methods use multiple machine learning models to obtain a predictive performance

better than any of its individual constituent members could have produced. Bagging is an ensemble method that uses random resampling of a dataset to construct models [13]. In classification scenarios, the random resampling procedure in bagging induces some classification margin, i.e., the gap between the classes, over the dataset. Additionally, when bagging is performed in different feature subspaces, the resulting classification margins are likely to be diverse, which is essential for an ensemble to be accurate. The methods take into account the diversity of classification margins in feature subspaces for improving the performance of bagging. First, it studies the average error rate of bagging, converts the task into an optimization problem for determining some weights for feature subspaces. Then, it assigns the weights to the subspaces via a randomized technique in classifier construction. Experimental results demonstrate that the ensemble method is robust for classification of noisy data and often generates improved predictions than any single classifier [14, 15].

In addition to their many other advantages, multiple-classifier systems hold the promise of developing learning methods that are robust in the presence of imperfections in the data in terms of missing features and noise in both the class labels and the features. Noisy training data tend to increase the variance in the results produced by a given classifier; however, by learning a committee of hypotheses and combining their decisions, this variance can be reduced. In particular, variance-reducing methods such as Bagging have been shown to be outstanding in the presence of fairly high levels of noise. In this study, the BDT classification is implemented using MATLAB [16].

### *Model Validation Methods*

Machine learning models are data driven and therefore resist analytical or theoretical validation. The models are constructed from an initial random state to a trained state using the training data sets and must be tested or validated using a different data set. Several validation approaches are available. Among them, the very popular one frequently used by researches is cross-validation.

In cross-validation, a series of SVM or BDT models are constructed, each time by dropping a different part of the data from the training set and applying the resulting model to predict the target. The merged series of predictions for dropped or test data are checked for accuracy against the observation. In one version of the cross-validation, called group cross-validation approach, data are divided into N groups. A total of N models are then constructed one by one using N-1 data groups for model training, and the remaining one group is for testing. Normally, N is chosen as 3, 5, and 10. At the end of this procedure, N predictions assembled from the dropped cases are compared with the observed targets to compute validation of model error for the cross-validation result. This cross-validation process was repeated N times, allowing each subset to serve once as the test data set. Ten-fold cross-validation is used in this investigation.

A number of methods are available to evaluate performance of binary classifiers. For a classifier with any given discrimination threshold, the number of cases correctly and incorrectly classified can be computed. This gives a confusion matrix with four numbers as shown in Table 1. TP is the number of true positives, i.e., how many cases are estimated by classifier as "Yes" events which actually are "Yes" events. Similarly we can define TN as the number of true negatives, FP as the number of false positives and FN as the number of false negatives. Using the statistics generated in Table 1, some frequently used classifier performance evaluation methods are described briefly below. More information about these methods can be found in Ref [17-19].

**Table 1 Confusion matrix for dichotomous ("Yes"/"No") events.**

| | | **Classifier Estimate** | |
|---|---|---|---|
| | | **Yes** | **No** |
| **Actual Observation** | **Yes** | TP | FN |
| | **No** | FP | TN |

The true positive rate (*TPR*) is the proportion of "Yes" observed events that were correctly estimated. *TPR= TP / (TP + FN).* It has a range of 0 to 1. If *FN*= 0, then the score goes to 1, which is the best value possible. The Overall Accuracy Rate (OAR) is defined as *OAR= (TP+TN) / (TP + FN + FP + TN).* It has a range of 0 to 1. "1" is the best classification performance score.

The false positive rate (*FPR*), which also called type I error rate, is the proportion of "No" observed events that were not correctly estimated. *FPR = FP /*

*(FP + TN)*. Its values also range from 0 to 1. If *FP*= 0, then the score goes to 0, the best one can expect. The false negative rate (*FNR*), also called type II error rate, is defined as *FNR = FN / (FN + TP) = 1-TPR*.

The Critical Success Index (*CSI)* is the proportion of true positives that were either estimated or observed. *CSI = TP / (TP + FP + FN)*. Its values range from 0 to 1 with a value of 1 indicating a perfect classification performance score. The *CSI* is more complete measure than TPR, FPR, or FNR. It depends on both false positives and false negatives, namely the *CSI* is sensitive to both type I and type II error rates.

Receiver Operating Characteristic (ROC) or simply ROC curve analysis has gained substantial popularity in the machine learning community lately [20-22]. A ROC curve is a graphical plot of the true positive rate, *TPR*, vs. false positive rate, *FPR*, for a binary classifier as its discrimination threshold varies. AUC (Area Under the ROC Curve) represents a ranking-based measure of classification performance. Its value can be interpreted as the probability of how well a classifier is able to distinguish a randomly chosen "Yes" example from a randomly chosen "No" example. In contrast to many alternative performance measures, AUC is invariant to relative class distributions, and class-specific error costs. For this reason, AUC is a commonly used performance metric for dealing with imbalanced data.

In general, the ROC curve bends toward the upper left corner where *TPR* are larger than FPR, and the AUC is then greater than 0.5. Where the curve lies close to the diagonal, the classification system does not provide any useful information, and AUC is approximately 0.5. The larger the AUC value, the easier the classifier can discriminate between a pair of positive and negative examples, so as to produce the better performance for the classifier.

To compare the classifier performance of SVM and BDT, OAR, CSI, and AUC classifier performance measures are used in this research.

## III. Experimental Data Setup

The weather impacted airport primary runway configuration selections were studied using the nonlinear binary classification models, SVM and BDT. The inputs for the models are airport terminal METAR weather data and the classification targets are different runway configurations. The models were trained and tested using tenfold cross-validation. The classification performances were evaluated using three metrics: OAR, CSI, and AUC.

During the analysis of airport AAR predictions, a threshold was designated by comparing the GDP AAR distribution with all AAR distributions and by referencing FAA airport capacity benchmark [23]. The threshold, called GDP AAR threshold in this paper, is used to group the data into two classes for the airport. For the AAR above the threshold, the airport capacity is optimum under good weather conditions. Otherwise, the capacity is reduced under adverse weather conditions.

The AAR prediction targets, AAR data at 2-hour, 4-hour, and 6-hour look ahead times were grouped into two classes of "Yes" and "No". AAR is denoted as "Yes" if its value is less than or equal to the GDP AAR threshold, "No" otherwise. The input data for AAR predictions include airport runway configuration information, current AAR, METAR weather, and T-WITI-FA weather forecast data. Applying both input data and 2 to 6 hour look ahead AAR classes, the SVM and BDT models were trained and tested using tenfold cross-validation. The classification performances were evaluated by the three classification performance metrics.

In this analysis, the data sources are the FAA National Traffic Management Log (NTML) database, the airport surface Terminal Forecast Weather Impacted Traffic Index, T-WITI-FA [24, 25], and the FAA Aviation System Performance Metrics (ASPM) database. All data over the years 2007 through 2009 were derived from these data sources.

### *GDP AAR Threshold*

The AARs over GDP events were selected and calculated from ASPM database using the GDP event start time and actual end time obtained from NTML database for each selected airport. Based on a comparison of airport GDP AAR, all AAR distributions for the airport and the airport capacity operation benchmark information, a GDP AAR threshold can be determined for that airport.

### *Airport AAR Data*

Observed airport hourly AAR data are collected from the ASPM database. For AAR

predictions, the current AARs are used as inputs and the AAR for two hour, four hour, and six hour look-ahead times are used as targets. The AAR numerical values of these targets are converted into a categorical attribute of "Yes" or "No" by the airport GDP AAR thresholds.

*Current Airport Terminal Weather Data*

Current terminal weather at airport is an important contributor to airport operations and planning. Actual hourly airport surface weather observations (METAR), such as wind, ceiling, visibility, and meteorological condition flags, were selected from ASPM database. These data were preprocessed to convert character records to numerical values and the missing data were filtered out. The processed METAR data were used as inputs for airport runway configuration selections and AAR predictions.

*Forecast Weather Data*

The forecast airport Terminal Weather Impacted Traffic Index, T-WITI-FA is provided by Alexander Klein from Air Traffic Analysis, Inc. It was computed based on airport Terminal Area Forecast (TAF) data, Collaborative Convective Forecast Product (CCFP) data and other air traffic information. The computed hourly data include 2-hour, 4-hour, and 6-hour forecast WITI data. Each forecast consists of seven factors. They are en-route convective WITI, local convective WITI, wind WITI, snow WITI, IMC WITI, volume/ripple effects WITI, and other WITI factor values. These seven factors and the sum of them for each forecast time were applied as inputs for AAR predictions. More details of these factors can be found in ref. 26.

*Airport Runway Configuration Data*

Airport operation data for runway configuration are collected from the ASPM database. For runway configuration binary classification analysis, we converted the runway labels into categorical attributes and used them as the targets. For AAR predictions, these data were preprocessed to convert character runway configuration labels to numerical values and are used as inputs for machine learning BDT and SVM models.

## IV. Results

This section presents the analysis results of using classification techniques to determine airport runway configuration and predict AARs grouped by GDP AAR threshold for the following four major US airports: Newark Liberty International Airport, San Francisco International Airport, Chicago O'Hare International Airport, and Atlanta International Airport. These four are typical due to their high GDP event rate caused by inclement weather with different dominant weather cause factors. The runway configuration and AARs were studied using the data collected over the years 2007 through 2009, which contain more than 17000 samples after data preprocessing.

*Newark Liberty International Airport (EWR)*

Among major US airports, EWR has one of the highest GDP event rate for the years 2007-2009. During these three years, EWR airport was affected by GDP in about 50% of days. For these GDP events, the average GDP duration is about 9 hours and 52% of them are caused by strong winds [26].

For EWR airport, most arrival aircraft are on Runway 4R-22L, while most departure traffics are on 4L-22R. The Runway 11-29 is used more often either by smaller aircraft or in cases where strong crosswinds occur on the two main parallel runways.

There are about 15 operational runway configurations in EWR airport operation. Among them, only five are primary runway configurations that are used at least 3% of the time during a year [4]. The five primary runway configuration usages, denoted as arrival | departure runway configurations, are listed in Table 2. As an example, the wind directions for the top 2 frequently used runway configurations are shown in Figure 1.

| Runway Configuration (Arrival\|Departure) | Annual Operation Percentages |
|---|---|
| 22L\|22R | 41% |
| 4R\|4L | 27% |
| 11, 22L\|22R | 15% |
| 4R, 11\|4L | 8.5% |
| 22L\|22R, 29 | 3.0% |

Table 2, EWR Primary Runway Configuration

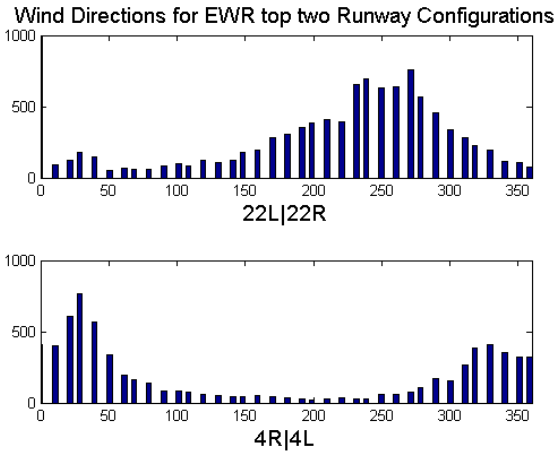


Figure 1, Wind direction distributions for EWR runway configurations with top two usages.

Table 3 lists the results of separating the most frequently used runway configuration "22L|22R" from the second frequently used, "4R|4L", and distinguishing "22L|22R" from all other runway configurations using SVM and BDT classifiers. The input to the models is airport METAR weather data. From this table, it is apparent that the BDT classification results are much better than SVM. Comparing the two columns in Table 3 with BDT classifier, the separation between "22L|22R" and "4R|4L" has an overall of 85% accuracy rate, while the separation between "22L|22R" and "all others" has only 76% overall accuracy rate. This may be explained as that some runway configurations, such as "11,22L|22R" or "22L|22R,29" included in the "all others" could be deployed under a similar weather condition for the runway configuration "22L|22R".

| | | 22L\|22R vs. 4R\|4L | 22L\|22R vs. All Others |
|---|---|---|---|
| SVM | OAR | 80% | 67% |
| | CSI | 0.74 | 0.41 |
| | AUC | 85% | 72% |
| BDT | OAR | 85% | 76% |
| | CSI | 0.78 | 0.54 |
| | AUC | 92% | 83% |

Table 3, EWR Runway configuration Estimation

By examining EWR AAR distributions for GDP and for all events in Figure 2 (a) and (b), the EWR GDP AAR threshold was determined as 40 arrival aircraft per hour. This 40 arrival aircraft per hour is also the EWR marginal rate [23]. Adopting this GDP AAR threshold of 40, the EWR AAR data in look-ahead times were grouped into two classes.

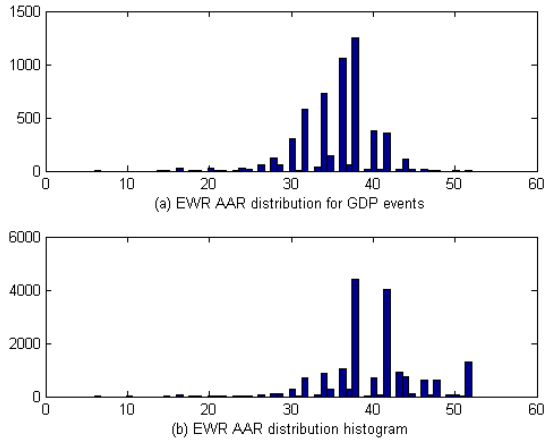


Figure 2, EWR AAR distributions

The AAR classification prediction results for 2-hour, 4-hour, and 6-hour look-ahead times are listed in Table 4. Once again, the BDT classification performance looks better than that from SVM. As an example, for 2-hour look-ahead AAR predictions, the AUC 95% confidence interval or two-sigma interval for BDT and SVM methods were computed as 92.7% to 93.2% and 81.8 to 82.6%, respectively. One can conclude that the AUC difference of 11% (93%-82%) between BDT and SVM is statistically significant as compared to their two-sigma intervals. Figure 3 shows that the ROC curve in solid for BDT classifier bends more toward the upper left cover, e.g., more far away from the diagonal than the ROC curve in dotted line for SVM, which clearly signifies that the overall accuracy of BDT is much higher than that for SVM. In other words, BDT has an ability to correctly classify the underlying subjects into their relevant subgroups better than SVM in this case. In terms of different look-ahead times, one can see from the table that the BDT classifier is doing very well for the 2 to 4-hour look-ahead AAR predictions while its performance measure of AUC for the 6-hour prediction is 86%; not as good as that for the 2 to 4-hour cases, but not bad, either.

| | SVM | | | BDT | | |
|---|---|---|---|---|---|---|
| | OAR | CSI | AUC | OAR | CSI | AUC |
| 2-h | 75% | 0.60 | 82% | 87% | 0.77 | 93% |
| 4-h | 71% | 0.54 | 78% | 81% | 0.68 | 89% |
| 6-h | 69% | 0.52 | 74% | 78% | 0.63 | 86% |

Table4, EWR AAR 2, 4, and 6 hour Prediction

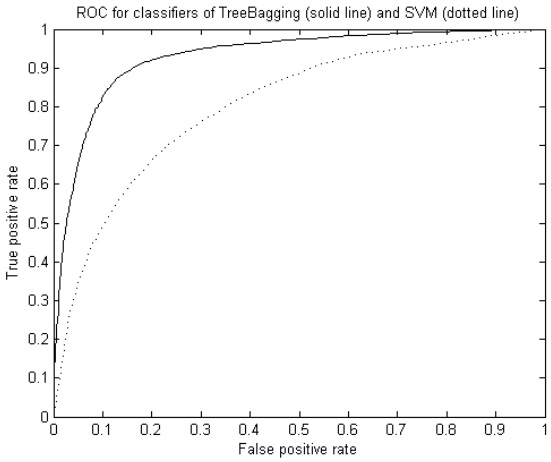


Figure 3, ROC curves of SVM (dotted) and BDT (solid line) for EWR AAR 2-hour prediction

*San Francisco International Airport (SFO)*

SFO has the second highest GDP events rate among major US airports. However, these events last only about 4.5 hours on average. About 88% of GDPs at SFO are caused by low ceilings due to marine stratus [26].

The current runway configuration at SFO consists of the following four: 10L-28R, 10R-28L, 1R-19L, and 1L-19R. There are more than ten operational runway configurations in SFO airport daily operation. Among them, the five primary runway configurations are listed in Table 5.

| Runway Configuration | Annual Operation Percentages |
|---|---|
| 28L,28R\|1L,1R | 61% |
| 28R\|1L,1R | 15% |
| 28L,28R \|28L,28R | 13% |
| 28L\|01L,01R | 3.2% |
| 19L,19R\|10L,10R | 3.0% |

Table 5, SFO Primary Runway Configuration

Table 6 lists the two SFO runway configuration selection results by SVM and BDT classifiers, respectively. The left column shows the results of distinguishing the top operational usage "28L, 28R|1L,1R" from all others. The right column displays the classification measures of separating the second runway configuration "28R|1L,1R" from the third one "28L,28R|28L,28R". Here it too reveals that the BDT classification results are superb comparing with the SVM. Both SVM and BDT classifiers illustrate that the classification performance for separating the two dissimilar runway configurations, i.e., the second and the third runway configurations counting from top, is much better than that for separating one from all other runway configurations.

| | | 28L,28R\|1L,1R vs. all others | 28R\|1L,1R vs. 28L,28R \|28L,28R |
|---|---|---|---|
| SVM | OAR | 68% | 81% |
| | CSI | 0.64 | 0.72 |
| | AUC | 69% | 86% |
| BDT | OAR | 73% | 86% |
| | CSI | 0.66 | 0.78 |
| | AUC | 77% | 92% |

Table 6, SFO Runway configuration Estimation

The AAR distributions for GDP and all events for SFO are shown in Figure 4 (a) and (b), respectively. A value of 40 was chosen as the SFO GDP AAR threshold. The number is also corresponding to the SFO marginal arrival rate [23]. The AAR classification prediction results for 2-hour, 4-hour, and 6-hour look-ahead times are listed in Table 7. Consistent with the previous findings, it proved again that the BDT classification works much better than SVM. The BDT classification measures with AUC for both 2 and 4-hour look-ahead cases are above 90%, which is excellent.

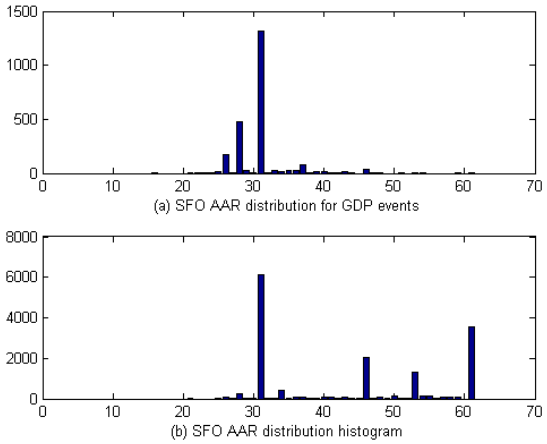


Figure 4, SFO AAR distributions

| | SVM | | | BDT | | |
|---|---|---|---|---|---|---|
| | OAR | CSI | AUC | OAR | CSI | AUC |
| 2-h | 83% | 0.71 | 91% | 88% | 0.78 | 95% |
| 4-h | 75% | 0.61 | 84% | 82% | 0.70 | 90% |
| 6-h | 72% | 0.55 | 79% | 79% | 0.65 | 87% |

Table7, SFO AAR 2, 4, and 6 hour Prediction

*Chicago O'Hare International Airport (ORD)*

ORD airport has high GDP events rate and these events last about 8 hours. About 29% ORD GDP events are caused by wind, 25% of them are by low ceilings, 15% by thunder storms, 14% by snow/ice, 8% by low visibility, and 6% by rain [26].

The current runway configuration at ORD consists of seven runways. There are more than 40 operational runway configurations used in ORD airport operation. All ten primary runway configurations are listed in Table 8.

| Runway Configuration | Annual Operation Percentages |
|---|---|
| 4R,9R,10\|4L,9R,10,32L,32R | 8.8% |
| 4R,9R,10\|4L,9R,32L | 7.5% |
| 27L,27R,28 \|22L,28,32L | 7.1% |
| 22R,27L,28 \|22L,32L,32R | 7.0% |
| 4R,9L,9R\|4L,9L,32L,32R | 6.5% |
| 22R,27L,27R\|22L,32L,32R | 6.2% |
| 4R,9R,10\|4L,9R,32L,32R | 3.9% |
| 22R,27L,28 \|22L,32L | 3.6% |
| 4R,9R,10\|4L,9L,32L | 3.4% |
| 27L,27R,28 \|22L,28,32L,32R | 3.3% |

Table 8, ORD Primary Runway Configuration

The runway configuration selections were estimated to distinguish the top runway configuration (i.e., "4R,9R,10|4L,9R,10,32L,32R") from the 2nd and the 3rd runway configurations counting from the top, respectively. The outcomes are listed in Table 9. The runway configurations of the top one and 2nd from the top are similar; the top and the third one from the top are not. Both SVM and BDT methods show that the classification performance metrics for separation of the two dissimilar runway configurations are significant better than that for separating the two similar runway configurations.

| | | 4R,9R,10\| 4L,9R,10,32L,32R vs. 4R,9L,10\| 4L,9R,32L | 4R,9R,10\| 4L,9R,10,32L,32R vs. 27L,27R,28\| 22L,28,32L |
|---|---|---|---|
| SVM | OAR | 65% | 85% |
| | CSI | 0.55 | 0.76 |
| | AUC | 69% | 92% |
| BDT | OAR | 76% | 91% |
| | CSI | 0.65 | 0.85 |
| | AUC | 85% | 97% |

Table 9, ORD Runway configuration Estimation

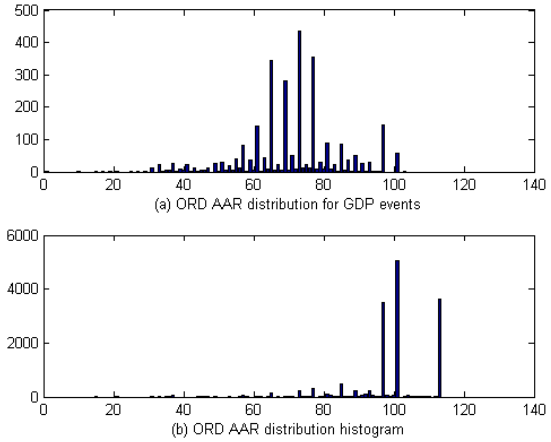


Figure 5, ORD AAR distributions

The ORD AAR distributions for GDP and all events are shown in Figure 5 (a) and (b), respectively. AAR of 95 was selected for the ORD AAR threshold. The number is close to the ORD marginal arrival rate listed in ref. 23. The AAR classification prediction results for 2-hour, 4-hour, and 6-hour look-ahead times are listed in Table 10. More same as before, the BDT results are superior than SVM. However, the ORD AAR classifier prediction performances are better in comparison to EWR or SFO airport. The overall accuracy rate of BDT is greater than 85%; the AUC is above 90%, even for the six hour prediction case.

| | SVM | | | BDT | | |
|---|---|---|---|---|---|---|
| | OAR | CSI | AUC | OAR | CSI | AUC |
| 2-h | 83% | 0.53 | 87% | 91% | 0.73 | 95% |
| 4-h | 79% | 0.47 | 83% | 89% | 0.65 | 93% |
| 6-h | 77% | 0.43 | 80% | 87% | 0.61 | 92% |

Table 10, ORD AAR 2, 4, and 6 hour Prediction

*Atlanta International Airport (ATL)*

ATL also has the relatively high GDP events rate. These GDP events normally last more than 6 hours. For these GDP events, 45% are caused by thunder storms.

The current runway configuration at ATL consists of five runways and the fifth runway, 10-28, opened at May, 2006. There are more than 10 operational runway configurations, among them only three, as shown in Table 11, are considered as primary runway configurations.

| Runway Configuration | Annual Operation Percentages |
|---|---|
| 26R,27L,28\|26L,27R | 49% |
| 8L,9R,10\|8R,9L | 35% |
| 26R,27L,28\|26L,27R,28 | 3.5% |

Table 11, ATL Primary Runway Configuration

The analysis of runway configuration selection was performed to separate the runway configuration of "26R,27L,28|26L,27R" from "8L,9R,10|8R,9L" and to distinguish "26R,27L,28|26L,27R" from all others. These results are listed in Table 12.

| | | 26R,27L,28\|26L,27R vs. 8L,9R,10\|8R,9L | 26R,27L,28\|26L,27R vs. all others |
|---|---|---|---|
| SVM | OAR | 85% | 78% |
| | CSI | 0.68 | 0.65 |
| | AUC | 91% | 84% |
| BDT | OAR | 89% | 81% |
| | CSI | 0.76 | 0.68 |
| | AUC | 95% | 88% |

Table 12, ATL Runway configuration Estimation

The ATL AAR distributions for GDP and all events are shown in Figure 6 (a) and (b), respectively. AAR of 105 was selected as the ATL AAR threshold. The AAR classification prediction results for 2-hour, 4-hour, and 6-hour look-ahead time are listed in Table 13. Once again, the findings are the same, i.e., the BDT classification performs much better than SVM. The BDT accuracy is quite good with AUC, which is above 90% in all cases including for 6-hour look-ahead time instance.

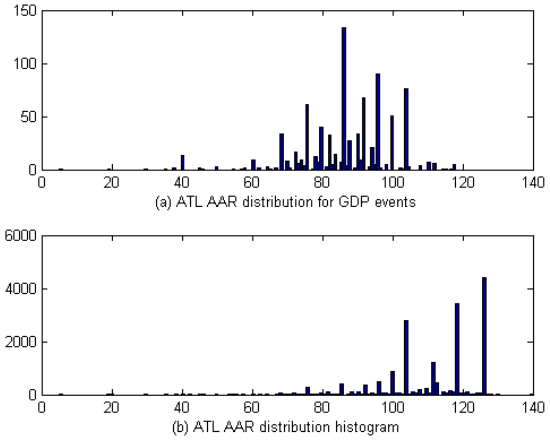


Figure 6 ATL AAR distributions

| | SVM | | | BDT | | |
|---|---|---|---|---|---|---|
| | OAR | CSI | AUC | OAR | CSI | AUC |
| 2-h | 82% | 0.58 | 89% | 90% | 0.76 | 95% |
| 4-h | 80% | 0.53 | 84% | 86% | 0.68 | 93% |
| 6-h | 78% | 0.50 | 82% | 84% | 0.64 | 91% |

Table13, ATL AAR 2, 4, and 6 hour Prediction

As a summary, the following observations can be made from results in all Tables for these four airports:

(a) Ensemble BDT consistently outperforms the single SVM classifier in all three classifier performance measures by statistically significant amounts during ten-fold cross validation testing.

(b) The BDT classifier provides very good estimates of the runway configuration based on the airport weather to distinguish dissimilar runway configurations. In such cases, overall accuracy rate is above 85%, CSI is greater than 0.75, and the most important measure, AUC, is above 90%. However BDT only offers fairly good classification results for distinguishing a runway configuration from similar or all other runway configurations, but the AUCs are still above 80%.

(c) The AAR classification predictions by BDT for 2 and 4 hour look-ahead times are excellent with an above 80% of overall accuracy and above 90% of AUC. For 6-hour AAR prediction, the performance of the BDT classifier is not bad, AUC is above 85%.

(d) The AAR prediction results using BDT models for EWR and SFO are not as good as for ORD and ATL. One reason could be that the dominant weather causes of GDP events for these four airports are different. The situations of fast changes in the wind direction and speed at EWR and rapid fog burn-off time at SFO in west coast may add more uncertainty on the AAR predictions.

## V. Concluding Remarks

This paper presents studies of the weather impacts on airport runway configuration selection and airport capacity using machine learning

approaches. It described how ensemble multiple classifier BDT model and traditional SVM can be used to estimate runway configuration selection and AAR 2 to 6-hour look-ahead predictions. The models evaluation was accomplished by ten-fold cross-validation. The performance of these two classifiers was determined by the overall accuracy rate (OAR), critical success index (CSI), and area under the ROC curve (AUC). The analysis, estimation, and prediction were achieved by using airport terminal METAR weather data, T-WITI-FA forecast data, airport runway configuration, and AAR information over the years 2007-2009.

The experimental results show that the proposed ensemble BDT classifier outperforms single SVM. Even though there is clearly room for fine-tuning and improving each of the algorithms, this conclusion should remain unchanged.

Since this analysis focused on weather impact on runway configuration selection and AAR predictions, other factors affecting runway configuration selections and AARs were ignored. For example, the noise abatement procedure information is not used for the runway configuration study. These factors would inject more noise and data imperfections into the analysis. The fact of BDT having better classification performance for our data demonstrates that multiple classifier systems are more robust in the presence of noise and other imperfections in data as compared to a single classifier system.

The BDT classifier performs well in both runway configuration selections and AAR prediction studies. This method is recommended as a decision support model in runway configuration selection and AAR planning of GDP events for TFM and airport daily operations.

## References


[1] Pace, Dave, 2009, ATM-Weather Integration Plan Overview, FAA AJP-B, Aviation Weather Office.

[2] FAA, 2009, Traffic Flow Management in the National Airspace System, FAA.

[3] Hoffman, Robert, A. Mukherjee, T. Vossen, C. Barnhart, B. Smith, 2001, Air traffic flow management, Quantitative Problem Solving Methods in the Airline Industry: A Modeling Methodology Handbook, Springer, Norwell, MA.

[4] FAA, Order JO 7210.3W, February 2010, Subject: Facility Operation and Administration.

[5] Neufville, Richard de, A. Odoni, 2003, Airport Systems: Planning, Design, and Management, McGraw-Hill.

[6] Lucic, Panta, M. Ohsfeldt, M. Rodgers, A. Klein, 2007, Airport runway capacity model review, Research report by CSSI and Air Traffic Analysis, Inc. for FAA ATO-P Performance Analysis and Strategy.

[7] Vapnik, Vladimir, 1998, Statistical Learning Theory, John Wiley and Sons, Inc., New York.

[8] Cortes, Corinna, Vapnik, V., 1995, "Support-Vector Networks", Machine Learning, 20.

[9] Wilks, Daniel S., 1995, Statistical Methods in the Atmospheric Science, Academic Press, pp. 467.

[10] Smith, David A., L. Sherry, G. Donohue, 2008, Decision Support Tool for Predicting Aircraft Arrival Rates, Ground Delay Programs, and Airport Delays from Weather Forecasts, Proceedings International Conference on Research in Air Transportation (ICRAT-2008), Fairfax, VA.

[11] Scholkopf, Bernhard, A. J. Smola, 2002, Learning with kernels, support vector machines, regularization, optimization and beyond, Cambridge, MIT Press.

[12] Chang, Chih-Chung, Chih-Jen Lin, 2011, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.

[13] Breiman, Leo, 1996, Bagging Predictors, Machine Learning, vol. 24, no. 2, pp. 123-140.

[14] Dietterich, Thomas G., 2000, Ensemble Methods in Machine Learning, Proc. Conf. Multiple Classifier Systems, pp. 1-15.

[15] Melville, Prem, N. Shah, L. Mihalkova, R.J. Mooney, 2004, Experiments with Ensembles with Missing and Noisy Data, Proc Fifth Int'l Workshop Multiple Classifier Systems, pp. 293-302.

[16] MATLAB R2011a, 2011, Statistics Toolbox, TreeBagger Class.

[17] Doswell, Charles A., R. Davies-Jones, D. L. Keller, 1990, On summary measures of skill in rare event forecasting based on contingency tables, Weather and Forecast, **5**, pp. 576-585.

[18] Swets, John A., 1996, Signal detection theory and ROC analysis in psychology and diagnostics: collected paper, Hillsdale, NJ.
[19] Fogarty, James A., R. Baker, Hudson S., 2005, Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction, ACM International Conference Proceeding Series, Proceedings of Graphics Interface, Waterloo, Ontario, Canada.
[20] Lasko, Thomas A., J. Bhagwat, K. Zou, L. Ohno-Machado, 2005, The use of receiver operating characteristic curves in biomedical informatics, Journal of Biomedical Informatics, 38(5), pp. 404–415.
[21] Zou, Kelly H., A.J. O'Malley, L. Mauri, 2007, Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models, Circulation, 6; 115(5), pp. 654–657.
[22] Gonen, Mithat, 2007, Analyzing Receiver Operating Characteristic Curves Using SAS, SAS Press, ISBN 978-1-59994-298-1.
[23] FAA, 2004, Airport Capacity Benchmark Report 2004.
[24] Klein, Alexander, T. MacPhail, Etc., 2009, NAS Weather Index: Quantifying Impact of Actual and Forecast En-route and Surface Weather on Air Traffic, 89th AMS Annual Meeting, Phoenix, AZ.
[25] Klein, Alexander, S. Kavoussi, R. Lee, 2009, Weather Forecast Accuracy: Study of Impact on Airport Capacity and Estimation of Avoidable Costs, 8th USA/Europe Air Traffic Management Research and Development Seminar Napa, California, (ATM2009).
[26] Wang, Yao, D. Kulkarni, 2011, Modeling Weather Impact on Ground Delay Programs, NASA/TN-3961.