Transportation Research Part C 120 (2020) 102819

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Towards online prediction of safety-critical landing metrics in aviation using supervised machine learning

Tejas G. Puranik [*,1], Nicolas Rodriguez [2], Dimitri N. Mavris [3]

Georgia Institute of Technology, Atlanta, GA 30332, United States

ARTICLE INFO

Keywords:
Air transportation
Risk
Safety
Supervised learning
Random forest
Landing performance

ABSTRACT

In recent years, due to the increased availability of data and improvements in computing power, application of machine learning techniques to various aviation safety problems for identifying, isolating, and reducing risk has gained momentum. Data collected from on-board recorders in commercial aircraft contain thousands of parameters in the form of multivariate time-series (continuous, discrete, categorical, etc.) which are used to train the machine learning models. Among the phases of flight, approach and landing phases result in the most accidents and incidents. The performance and trajectory of the aircraft during the approach phase is an indicator of its landing performance which, in turn, affects incident or accident probability such as runway excursions. Landing performance is commonly measured using metrics such as landing airspeed, vertical speed, location of touchdown point on runway, etc. While current applications of machine learning to aviation focus on retrospective insights to implement corrective measures, they offer limited value for real-time risk identification or decision-making as they are inherently reactive in nature. In this work, a novel offline-online framework is developed for building a global predictive model offline to predict landing performance metrics online. The framework leverages flight data from the approach phase between certain approach altitudes (also called *gates*) in order to train the offline model to predict the landing true airspeed and ground speed using a Random Forest regression algorithm. Permutation importance is used to identify and retain important features among those available and the training data is balanced with respect to flight safety events to ensure a good performing model. The developed global model is robust and can predict landing true airspeed and ground speed with root mean square errors of 2.62 and 2.98 knots respectively over six different airframes operating at over seventy airports. The model operates fast in prediction mode which, coupled with the ability to provide the prediction at an altitude where a go-around decision may be made, makes it particularly suitable for online application. The framework is demonstrated using data obtained from commercial airline operations that contains thousands of flight records and performs better than existing techniques in literature at predicting true airspeed and ground speed at touchdown.

* Corresponding author.
E-mail address: tpuranik3@gatech.edu (T.G. Puranik).

[1] Research Engineer II, Aerospace Systems Design Laboratory, Georgia Institute of Technology, Atlanta, GA, United States.

[2] Graduate Research Assistant, Aerospace Systems Design Laboratory, Georgia Institute of Technology, Atlanta, GA, United States.

[3] S.P. Langley NIA Distinguished Regents Professor, Director of Aerospace Systems Design Laboratory, Georgia Institute of Technology, Atlanta, GA, United States.

https://doi.org/10.1016/j.trc.2020.102819
Received 14 April 2020; Received in revised form 15 August 2020; Accepted 23 September 2020
Available online 16 October 2020

1. Introduction

Over the past few decades, the aviation industry has witnessed a steady increase in the volume and frequency of air traffic worldwide. The safety record of aviation has also been steadily improving due to advancements in technology, focused efforts of government, manufacturers, and operators through various safety improvement programs, and increased awareness of risks (Statistical Summary of Commercial Jet Airplane Accidents - Boeing Commercial Airplanes, 2017). According to the Federal Aviation Administration (FAA), the demand for air travel and traffic is predicted to grow steadily over the next two decades at a rate of approximately 1.8% annually (Federal Aviaition Administration Aerospace Forecasts Fiscal Years, 2017). Commercial aviation operations are expected to double in volume and increase in complexity over this time period. It is therefore important to maintain and improve the safety record of aviation operations through this period of growth using innovative and advanced techniques.

In recent years, there has been a proliferation of data-driven approaches to risk assessment and safety improvement in the engineering domain (Hegde and Rokseth, 2020). Transport category airplanes (Part 121) have certain minimum requirements related to digital flight data recorder (DFDR) systems set by the FAA for U.S. carriers (Federal Aviation Administration, 2011). This includes a minimum set of more than ninety parameters related to the state, attitude, control surface deflections, engine information, environmental conditions, global positioning system (GPS) information, and others that need to be collected in-flight. In reality, modern commercial aircraft are equipped to record thousands of parameters at a high frequency throughout the duration of the flight (Campbell, 2020). This data recording capability coupled with the high volume of operations results in an explosion of flight data available for safety analysis to airliners and operators. Historically, accidents have been the primary triggers for identifying problems and developing mitigation strategies (Logan, 2008). The aviation industry has been shifting from a *reactive* to a *proactive* and *predictive* approach where potential unsafe events and precursors are identified beforehand and mitigation strategies are implemented to prevent loss of life. Traditional techniques of flight data analysis have focused on a continuous cycle of data collection from on-board recorders, retrospective analysis of flight-data records, identification of operational safety exceedances, design and implementation of corrective measures, and monitoring to assess their effectiveness. According to Campbell (2020), the number of operators using flight data collected on-board using DFDR has increased exponentially in the past few decades. Airliners have established flight safety divisions to analyze flight data, investigate safety issues, and proactively identify risks. Similarly, collaborations between industry and government such as the Aviation Safety Information Analysis and Sharing (ASIAS) (Federal Aviation Administration - Aviation Safety Information Analysis and Sharing (ASIAS), 2017) have spurred research in data-driven techniques for aviation safety.

The approach and landing phases are identified as some of the most critical phases of flight from the perspective of safety of operations (Sherry et al., 2013; Wang et al., 2016b). Errors or deviations from the flight path during approach and landing are likely to end in accidents or incidents as they have a small safety margin owing to the limited time for a pilot to correct the error or react to a deviation. Due to these reasons, the present work focuses on improvement of safety and risk assessment in the approach and landing phases of flight. The FAA and individual airlines publish various Stabilized Approach Criteria (SAC) in order to aid decision making in the approach phase. The FAA defines a stabilized approach as: "*A stabilized approach is characterized by a constant-angle, constant-rate of descent approach profile ending near the touchdown point, where the landing maneuver begins*" (Federal Aviation Administration Advisory Circular, 2003). A stabilized approach is viewed as one of the key features of safe approaches and landings in air carrier operations. SAC are widely used by airlines, although there is usually a variety in the parameters used in these criteria. Some airlines have as little as four parameters while others use significantly more. Previous research indicates that glideslope deviation, localizer deviation, rate of descent, and speed change, are some of the key factors in stable approaches (Flight Safety Foundation, 2000; Moriarty and Jarvis, 2014; Wang et al., 2015, 2016a,b; Sherry et al., 2013). Various mitigation techniques have been proposed for reducing the risk due to unstable approach such as flight envelope estimation techniques (Schuet et al., 2017). As seen from its definition, proper energy management is one of the key features of a stabilized approach. Energy state awareness and energy management are critical concepts in the characterization, detection, and prevention of safety-critical conditions (Puranik et al., 2017; Flight Safety Foundation, 2000). During the approach and landing phase in particular, many safety events are defined and monitored in relation to the state of the aircraft during various important decision gates such at 1000 feet above touchdown, 500 feet above touchdown, etc.

Appropriate decision-making by flight crews during the approach phase of flight is critical due to the limited amount of time available for corrective action. In addition, there are a number of constraints that pilots have to be aware of, such as Air Traffic Control (ATC) constraints, weather conditions, aircraft state, communication, etc. Thus, providing a reliable prediction of encountering risky situations in the future is valuable for online condition monitoring, particularly during the approach phase. Presently, airline procedures state that pilots make a go-around decision using some variants of the SAC at 1000 ft (for IMC conditions) or 500 ft (for VMC conditions) above the ground, or so-called decision gates (Campbell et al., 2018). If it is deemed unsafe to continue the landing, pilots perform a go-around and attempt the landing again to avoid risk of accidents like runway excursion or controlled flight into terrain (CFIT). Therefore, an accurate data-driven prediction of anticipated landing performance metrics (such as landing airspeed and ground speed) during the approach phase would be of great value to aid this decision.

Recently, machine learning techniques have been applied on several problems with a large degree of success. Three different types of machine learning techniques relevant in this context are supervised, semi-supervised, and unsupervised learning. Supervised learning relies on a labeled training set to build models offline that can be used with new, unseen data online to provide predictive capabilities. Semi-supervised techniques only require a set of training data that contains mostly nominal system behaviors which is used to train offline models and identify anomalies in test data. Finally, unsupervised techniques operate under the assumption that the given data set may contain normal and anomalous data in any proportion and separate this data into different clusters. Despite numerous applications in an unsupervised and semi-supervised context in the aviation safety domain, those in a supervised learning context are relatively scarce. On the other hand, tremendous advances have been made in various domains using supervised learning.

The availability of truth labels in supervised learning and the ability to obtain tangible validation metrics without the need for subject matter expert involvement make this application of ML techniques worth exploring in more detail. Therefore, in this this paper, a framework is developed for an online prediction model of critical landing parameters using supervised learning models to aid in decision-making and risk assessment during the approach and landing phases of flight.

The rest of paper is organized as follows: Section 2 provides an overview of existing applications of machine learning in air transportation system domain and their limitations and highlights the research objective and contributions of this work. Section 3 contains a detailed description of the offline/online prediction framework developed in this work, including the descriptions of the data processing and machine learning algorithm application. Section 4 shows the application of the developed methodology on real-world data collected during routine operations from an airline and the results. Section 5 provides concluding remarks and avenues of future work.

2. Background and research objective

The use of machine learning techniques for solving complex problems in the transportation domain has gained in popularity in recent years. Hegde and Rokseth (2020) have conducted a survey of such machine learning applications to engineering risk assessment. Numerous applications of machine learning in the aviation safety domain have recently been published in literature. These can typically be categorized based on the type of data (qualitative or quantitative) and the techniques utilized (supervised, semi-supervised or unsupervised) in analysis. One category of studies focus on improvement of safety using qualitative data in the aviation domain. These include studies such as system-level analysis of commercial aviation operations (Moriarty and Jarvis, 2014), identification of trigger events and high-risk chains in helicopter operations (Rao and Marais, 2015), retrospective analysis of approach stability (Rao and Puranik, 2018), time series forecasting of go-around incidents (Subramanian et al., 2018), among others. Many of these methods in literature forecast cumulative quantities such as sum of accidents or incidents in a year.

Another category of applications which is perhaps the most popular in aviation safety domain is anomaly detection using quantitative time-series data with semi-supervised or unsupervised techniques. In the data mining community, anomaly detection is defined as the *"task of obtaining patterns in data that do not conform to a well defined notion of normal behavior"* (Chandola et al., 2009). The objective of anomaly detection techniques in aviation is to detect abnormal flights within routine flight data without any prior knowledge of what constitutes an anomaly. In some cases, subject matter expert knowledge or physics-based assumptions may be used to augment unlabeled data with definitions of exceedances or safety events. Anomalous flights thus obtained (sometimes as ordered lists based on their 'anomalousness') are then further analyzed by experts for potentially dangerous/unsafe conditions. Basora et al. (2019) have provided a detailed review of anomaly detection techniques specifically applied to the aviation domain some of which are highlighted here.

The literature in aviation safety is mainly aimed at identifying two types of anomalies in the data – *Flight Level Anomalies* in which the entire flight record or phase of flight considered are anomalous, and *Instantaneous Anomalies* in which only an instant or small part (a few seconds) of the flight record is anomalous. There is a wealth of studies on identification of flight level anomalies using flight data. SequenceMiner (Budalakoti et al., 2009) is a software used to detect anomalies in discrete parameter sequences by learning from a model of normal switching. This technique detects flight-level anomalies but is limited to discrete data. Das et al. (2010) have developed Multiple Kernel Anomaly Detection (MKAD) which applies a one-class support vector machine for anomaly detection. MKAD identifies flight level anomalies well in data that contains discrete and continuous parameters. Li et al. (2015) have developed ClusterAD – an algorithm that uses density-based clustering for anomaly detection. Matthews et al. (2013) have discussed and summarized the aviation knowledge discovery pipeline using various state-of-the-art algorithms. Jarry et al. (2020) develop a method that leverages Functional Principal Component Analysis (FPCA) and HDBSCAN algorithms to detect atypical, or anomalous, approaches.. Previous work by the authors (Puranik and Mavris (2018)) focused on identifying anomalies using energy-based metrics and classification techniques in General Aviation (GA).

Techniques that are used to identify instantaneous anomalies are generally different than those used to identify flight-level anomalies. Amidan and Ferryman (2000) have utilized Singular Value Decomposition (SVD) to identify instantaneous anomalies. They mapped the five seconds before and after each recorded data point and used a regression model to identify outliers. Mugtussids (2000) has used Bayesian classification to distinguish between typical data points, that are present in the majority of flights, and unusual data points that can be only found in a few flights. Li et al. (2016) developed ClusterAD – Data Sample, which is a technique leveraging a mixture of Gaussian models to identify probability of a sample being anomalous during take off, approach, and landing. In previous work by the authors (Puranik and Mavris (2019)), the application of energy metrics as features along with classification techniques to identify instantaneous anomalies in GA has been demonstrated.

Compared to these unsupervised and semi-supervised techniques for anomaly detection, supervised techniques are relatively scarce. Inductive Monitoring System (IMS) (Iverson, 2004) is a technique that relies on a training set consisting of typical system behaviors which is compared with real-time data to detect anomalies. Each point is monitored standalone and therefore, the temporal aspect of anomalous sub-sequences is lost when identifying anomalies. Recently, Lee et al. (2020) have used Random Forest classification algorithm to identify precursors to three different aviation safety events using supervised learning. Tong et al. (2018) have used Long Short-Term Memory (LSTM) models to predict the future speed of an aircraft using historical data. Similarly, Zhang and Zhu (2018) and Tong et al. (2018) have predicted the probability of hard landings using time-series data and deep neural networks. In all three of these studies, the a single metric is chosen to be predicted at time intervals slightly beyond the present state of the aircraft. Despite high accuracy in predicting parameters at future time steps, due to the predictions being only a few seconds in the future, it provides limited real-time applicability for flight crews or air traffic controllers. Diallo (2012) presents results on landing speed

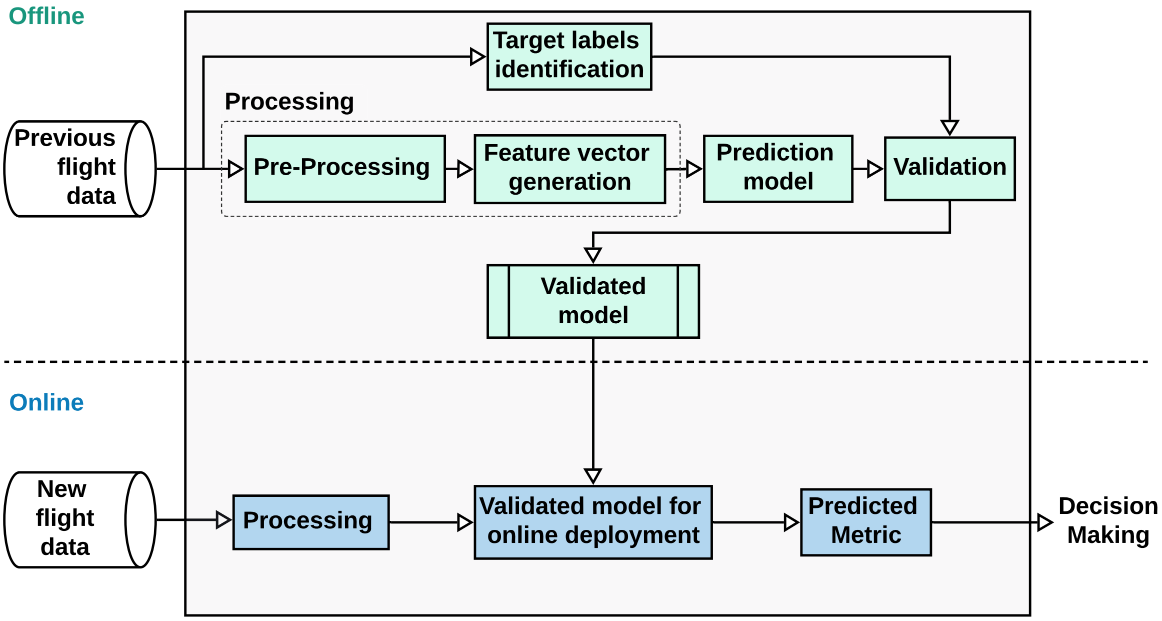


Fig. 1. Overview of the offline/online approach for landing metric prediction.

prediction using an Artificial Neural Network (ANN). However, the average error on the prediction is relatively high (12%) and there is limited discussion about the details of feature vector generation or sensitivity analysis of parameters is made. Jarry et al. (2020) have used LSTM networks to predict on-board parameters such as fuel flow rate, flap configuration, etc. using ground surveillance data. Martınez et al. (2019) have used boosting algorithms and LSTM models to forecast the likelihood of unstable approaches using past trajectory data. It is also noted that most of the applications in literature highlighted in this section have been demonstrated on a homogeneous fleet of aircraft typically at a single airport of operations. Therefore, the generalizability of the approaches is uncertain. While applications of supervised learning are scarce in the air transportation safety domain, it has been widely applied in other fields and therefore, it is of interest to be able to identify appropriate applications for supervised learning in the air transportation safety domain and to demonstrate their impact on real-world flight data. In consideration of the preceding observations, the research objective and contributions of this work are identified in Section 2.1.

2.1. Research objective

The overarching research objective of this work is stated as follows:

To demonstrate a novel supervised learning application for the online prediction of safety-critical landing metrics during approach phase in commercial aviation operations.

Some of the main limitations of existing supervised learning approaches were that they cannot provide predictions of critical quantities ahead in time to be useful for real-time decision making. Additionally, it is important to identify appropriate metrics of interest for the model to predict that correspond to safety margins and safe operation of aircraft. Some of the main contributions of this work are:

1. Provides a novel online predictive model of aircraft landing performance (landing true airspeed and ground speed) using data collected on-board an aircraft during the approach phase for multiple airframes at a variety of airports of operation
2. Demonstrates accurate prediction of the critical metrics further ahead in time than existing approaches and at a higher accuracy thereby unlocking the potential to aid in decision-making and real-time deployment
3. Introduces innovations in generating feature vectors and target labels using a flexible approach that can be easily replicated for other metrics of interest

In addition, it is noted that in order to be eventually useful for an online prediction perspective, the prediction time of any developed algorithm should be of the order of a few seconds as those are the time frames available for pilots, operators, and controllers to make such decisions.

3. Methodology

This section describes the methodology developed for the data-driven online predictive model generation in a general setting. There are two main components of the methodology. The first one is the offline (or training) component in which historical flight data

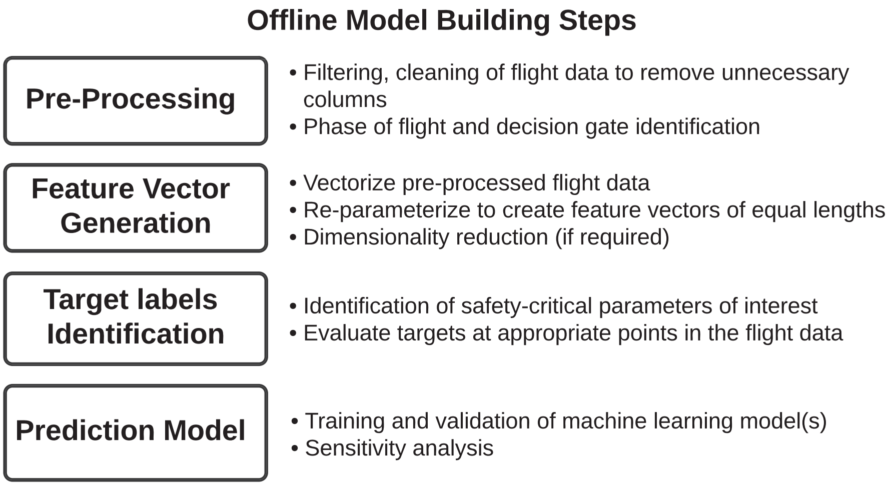

Offline Model Building Steps
Pre-Processing
• Filtering, cleaning of flight data to remove unnecessary columns
• Phase of flight and decision gate identification
Feature Vector Generation
• Vectorize pre-processed flight data
• Re-parameterize to create feature vectors of equal lengths
• Dimensionality reduction (if required)
Target labels Identification
• Identification of safety-critical parameters of interest
• Evaluate targets at appropriate points in the flight data
Prediction Model
• Training and validation of machine learning model(s)
• Sensitivity analysis

Fig. 2. Overview of the steps during the offline (training) phase for landing metric prediction.

is processed, analyzed, and used to train and store a database of models that can be used for prediction. The second component is the online (or test) phase in which pre-trained models are deployed to provide a prediction of critical metrics of interest using new flight data that has not been used in the offline phase. Fig. 1 provides an illustration of the different steps of the methodology in each of the two components. The various blocks in the general methodology are described in further detail in this section followed by the implementation of this methodology on real-world data for a particular landing metric in the results (Section 4). It is noted here that the *Pre-processing* and *Feature vector generation* steps in the offline phase together correspond to the *Processing* step in the online phase in Fig. 1.

3.1. *Description of flight data*

Machine learning techniques require the availability of large quantities of data. Aviation safety data can be obtained from a variety of sources such as Flight Data Recorders (FDR), Cockpit Voice Recorder (CVR), Aviation Safety Action Program (ASAP), Aviation Safety Reporting System (ASRS), Flight Operations Quality Assurance (FOQA), etc. Among these *FOQA* data consists of regularly recorded aircraft sensor measurements and switch settings. The data obtained from the flight data recorder are a multivariate time series whose lengths typically vary due to varying duration of each flight. The data collected consists of thousands of parameters (numerical, discrete, categorical, text, etc.) recorded at a frequency of up to 16 Hz. The parameters in the flight data can be divided into different categories and levels based on their source system/sub-system in the aircraft. Atmospheric data refers to data gathered from pitot tubes, barometers, thermometers, etc. It includes airspeed, wind speeds, pressure altitude, atmospheric temperature, etc. Attitude data refers to roll, pitch, yaw angles and their corresponding rates and accelerations. GPS data contains the latitude, longitude, altitude, and related rates. Engine data contains RPM, Exhaust Gas Temperatures (EGT), Cylinder Head Temperatures (CHT), oil temperature and pressure, fuel flow rates, fuel quantities, etc. Control data contains the deflection of flaps, elevator, aileron, rudder, etc. Communications data includes details about the communication status of the vehicle, such as the comm frequency. Navigation data includes information on any way-point guidance or autopilot features. These are among the numerous categories of parameters typically present in flight data. Typical FOQA programs involve a continuous cycle of data collection from on-board recorders, retrospective analysis of flight-data records, identification of operational safety exceedances, design and implementation of corrective measures, and monitoring to assess their effectiveness (Advisory Circular, 2004). The steps in the offline model building stage are described as outlined in Fig. 2 in the next few sub-sections.

3.2. *Pre-processing*

The FOQA data obtained needs to be processed and cleaned in order to be applicable in a machine learning context. Various operations are performed on each flight data record during the pre-processing step. The first step in this process is cleaning the flight data in order to remove irrelevant or empty parameters (columns) that may be present in the original data. Examples of this type of column are Raw METAR Text (unnecessary column) or columns containing all not-a-number (NaN) entries. Highly correlated or redundant columns are not removed in this step as they can prove to be valuable in scenarios such as sensor failures. The columns identified in this first step are removed from every flight data record in the available data. This step helps reduce the size of the data without losing any valuable information.

The second step of pre-processing is the identification of phases of flight. In the data available in the present work, different phases of flight are identified by discrete parameters existing in the FOQA data that signal the beginning and end of a particular phase. Therefore, the approach phase of flight as identified in the FOQA data is extracted for each flight data record in order to be analyzed.

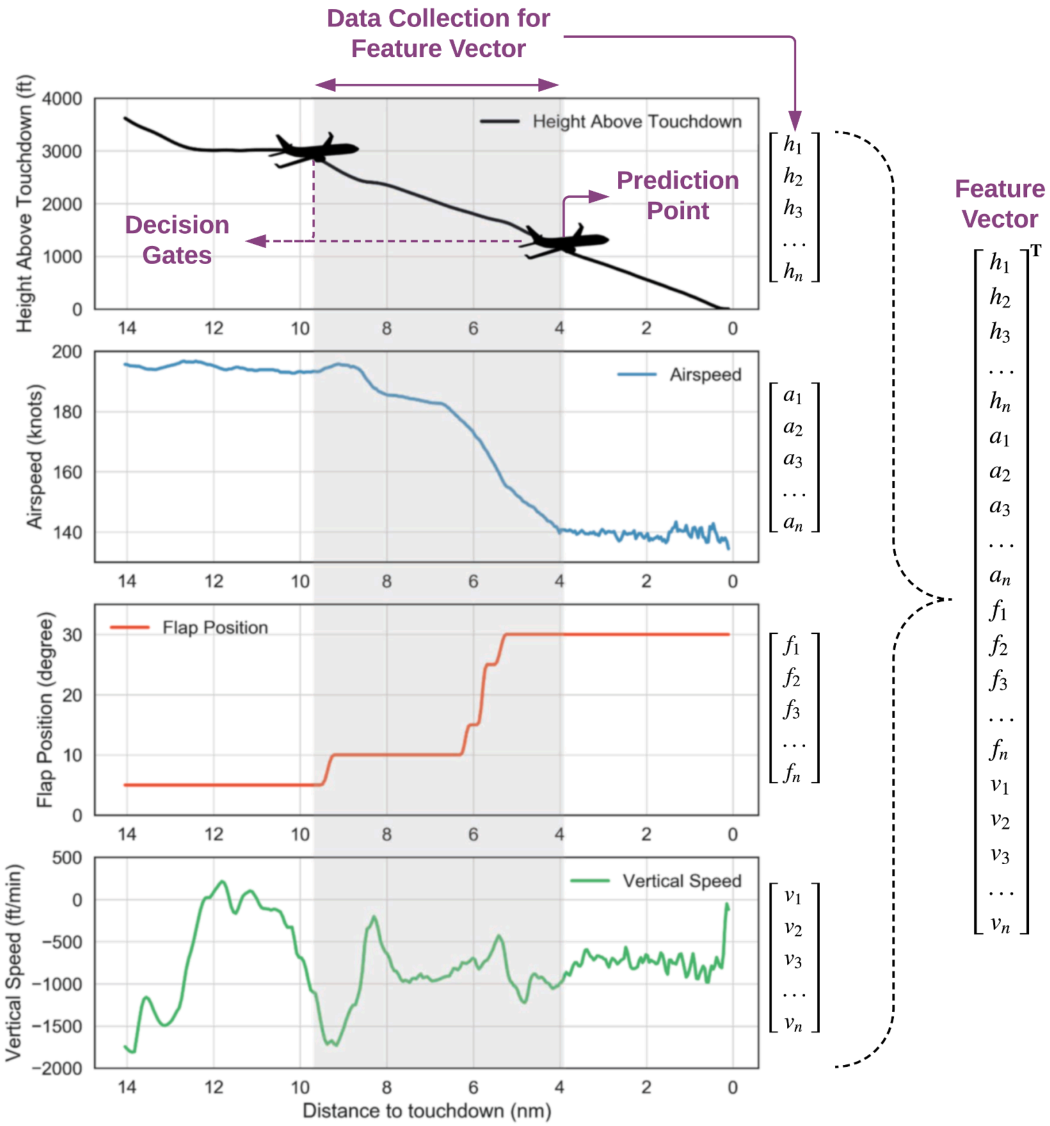


Fig. 3. Process of generating feature vectors for each flight by using recorded flight data between two decision gates shown for a notional flight.

Among the data available, this phase of flight begins anywhere between 6000 to 10000 feet above touchdown elevation and ends at the point when the aircraft first touches the runway. For scenarios in which approach phase of flight is not explicitly captured using parameters in the flight data record, a data-driven method such as that provided in previous work by the authors (Puranik et al., 2016) can be used where the touchdown point on the runway is estimated using recorded parameters such as Weight-on-Wheels (WoW) or altitude difference/vertical speed. The approach phase is then obtained by marching back in time from the touchdown point for a particular number of time steps for each flight data record.

3.3. *Feature vector generation*

One of the crucial steps in the framework is the generation of feature vectors for the machine learning model. A considerable amount of time and resources are typically spent in machine learning applications to identify the right features and extract them from the available cleaned data (sometimes referred to as *feature engineering*). The feature generation process developed for this work is outlined in Fig. 3 using a notional flight data record example. The pre-processed and extracted data from previous step contains recordings from the entire approach phase up to the touchdown point. However, this data has one major limitation from applicability in an online machine learning context – the duration of data from each flight is different and therefore cannot be directly compared against each other to build a consistent model. Additionally, if a prediction is to be made about landing performance that provides enough time for the pilot to react, it should be sufficiently before the airplane actually touches down on the runway. Thus, a specific subset of this data is extracted based on the knowledge of flight operations in the approach phase and the nature of the data.

As noted earlier, flight operators and airlines typically make decisions regarding the approach stability and safety at different points of time during the approach phase. If the approach is deemed unstable and potentially unrecoverable, pilots are advised to perform a go-around maneuver and attempt the landing again. Campbell et al. (2018) conducted human-in-the-loop simulator experiments and found that the go-around decision point should occur above the 100-ft gate. Considering these insights various decision gates are used in this work to make predictions of critical landing metrics. For the data collected to be comparable across different flights, all data between two different heights above touchdown is collected and reparameterized to have the same length using interpolation between the available values. The decision to quantize the data based on altitude bins rather than distance to landing can lead to limitations in certain cases, especially if the aircraft holds its altitude below 1000 ft for a significant amount of time. However, as the altitude is more readily accessible from an online prediction perspective and as many decisions are made at various altitude gates, in this paper, the altitude-based quantization is used. While the overall methodology is capable of being recast into a distance-to-landing based prediction method, that is outside of the scope of this work.

Fig. 3 shows the data generated for a notional flight. As an example, the data between 3000 feet and 1000 feet above touchdown (shown by the gray shaded region) is collected and reparameterized such that each parameter has the same predefined number of samples (denoted by n in the figure). In this work, the reparameterization is done based on altitude increments, for example, one sample every 20 feet. Once this data is collected and concatenated, the feature vector for each flight can be constructed as shown in the Fig. 3. In this manner, a unique feature vector for each flight is created and can be used to train the offline model for metric prediction. It is noted that the choice of decision gates can have an impact on the accuracy and usefulness of predictions and therefore, different decision gate combinations are explored in this work. The second decision gate shown in Fig. 3 is the prediction point for the algorithm. At this point, landing metrics of interest will be predicted using the trained offline model. Indeed, the closer the prediction point is to the actual touchdown, the more accurate the model will be. However, it will also provide less time for the pilot to react or correct their approach. Therefore, a trade-off exists between accuracy of the model and its usability in an online setting. As evidenced later in A, collecting data between 1000 and 300 feet rather than 3000 and 1000 feet (as shown in Fig. 3) provides the best model performance in the present context.

Despite reducing the number of parameters after pre-processing, hundreds of parameters in each flight data record are still retained at this step. Moreover, each parameter has multiple samples within the two gates under consideration. For example, between 3000 and 1000 feet altitudes, if data is collected from 600 sensors every 20 feet, each flight record would be represented with 60,000 features. However, from a flight safety and predictive modeling perspective all of these may not be equally important. Therefore, down-selection of the most influential parameters is vital to ensure that a robust and scalable model is built. This is achieved using feature selection. By performing feature selection, data or parameters that are irrelevant are ignored which helps improve accuracy and computational costs during online prediction phase. There are typically three major methods of feature selection – wrappers, filters, and embedded methods (Guyon and Elisseeff, 2003). Wrappers utilize the model of interest as a black box to score subsets of variables according to their predictive power. Filters select subsets of variables as a pre-processing step, independently of the chosen predictor. Embedded methods perform variable selection in the process of training and are usually specific to given machine learning models. Since this work aims to build a methodology for offline predictive model without particular emphasis on which specific model gets used, wrapper methods are best suited as it lends itself to the use of off-the-shelf machine learning models. In this work, feature selection is performed using a wrapper method called permutation importance introduced by Breiman (2001) for random forest models. It computes the drop in out-of-bag performance after permuting the values of a feature. For further details, readers are encouraged to consult Breiman (2001).

3.4. Target labels identification

In supervised learning (either classification or regression) target labels are used in the offline model training phase which represent ground truth. In many applications the choice of labels is self-evident from the task. For example, in problems like image classification the label could be a binary variable indicating whether the model has correctly identified the image class. In regression problems, typically the label is the true value of the output being predicted. One of the main challenges with the application of supervised learning in the aviation safety domain in particular is the identification of appropriate targets to predict. For example, in previous applications, flight data from the approach phase is used to classify normal and anomalous flights. However, since there is no ground truth data available for this classification unsupervised techniques are preferred and limited validation is performed.

Safety is a concept that may be difficult to quantify, because it is associated with the absence of something rather than its presence (Reason, 2000). However it may be approximated using metrics calculated from available data. Previous studies have shown that improper or poor energy management and loss of energy state awareness (LESA) are among the top contributors to Loss of accidents in aviation (Belcastro and Jacobson, 2010). Energy state awareness and energy management are critical concepts in the characterization, detection, and prevention of safety-critical conditions. Therefore, energy-based metrics (Puranik et al., 2017) such as those that characterize the energy state and safety boundary conditions of the aircraft hold significant potential for improving operational safety because they explicitly address poor energy management and state awareness as the top contributing factors underlying safety events.

Recent studies by Campbell et al. (2018) found that at all decision gates deviation in reference velocity had the strongest effect on touchdown performance. A runway excursion[4] is defined as "*An event in which an aircraft veers off or overruns the runway surface during either take-off or landing*". Runway excursions lead to more runway accidents than all the other causes combined. During approach and landing, unstable approaches are one of the main causes of runway excursions (due to excessive speed), hard landings, long landings, etc. Being able to predict the speed at which an aircraft is likely to touchdown can provide valuable insight into the risk of runway excursion and can help with go-around decision-making. Therefore, in this work, the speed of the aircraft at touchdown (in the form of two metrics: true airspeed and ground speed) is predicted as a safety metric in the supervised learning model.

The touchdown point is identified using a parameter available in the recorded data. This parameter is zero when the aircraft is airborne and changes to one when it touches the ground. To account for sensor noise and fluctuations, the touchdown speed is averaged using two samples before and after the touchdown point parameter changes to one and remains at that value. In this manner, a single target label is obtained for each flight in the training and test data set for offline model construction. While touchdown true airspeed and ground speed are used as the main target for prediction in this work, it is noted that any number of other metrics relevant to safety can be substituted for this metric in the developed framework.

3.5. Prediction model

The final part of the offline model building consists of training the model using the feature vectors identified and the target labels chosen. There are numerous choices for prediction models in the time series supervised regression problem. Among literature on similar methods for prediction of aviation safety metrics Diallo (2012) use a Neural Network to predict airspeed at touchdown. Tong et al. (2018) use a long short-term memory (LSTM) Neural Network to predict metrics at a time *t* using data at a time *t-1*. Li (2010) compares different Multilayer Perceptron models for the prediction of fuel consumption. Lee et al. (2014) use a Random Forest algorithm for the fault detection of aircraft systems. Petukhova et al. (2018) have demonstrated the superior performance of using a Random Forest algorithm for time series regression compared to auto-regressive methods. While the choice of a machine learning model is important, the aim of this study is not to compare different machine learning algorithms but rather to demonstrate the value and improvement of the developed framework over existing approaches. Due to its simplicity, accuracy, and ease of tuning (Breiman, 2001; Petukhova et al., 2018), the Random Forest Regression algorithm has been chosen for building the prediction model in this framework. Random forest has many advantages; it typically requires lower number of hyperparameters to tune than deep networks, it works well with heterogeneous datasets containing continuous, discrete, and categorical features, it has lower risk of overfitting, etc. A brief description of random forest regression is provided here and the readers are directed to Breiman (2001) for the detailed theoretical framework and scikit-learn documentation[5] for specifics on the implementation.

A random forest is a meta estimator that is a combination of tree predictors. A tree predictor is a supervised machine learning algorithm used for both classification and regression. At each node of the tree, the dataset is split based on a feature that maximizes the homogeneity within each group. A forest is built with a random group of uncorrelated trees and the final prediction is an average of each predicted target from each tree. In this work, the hyperparameters of the Random Forest are tuned using a combination of grid-search and Design of Experiments (DOE) such as those proposed by Lujan-Moreno et al. (2018) along with using the k-fold cross validation technique (Unpingco, 2016) for each hyperparameter combination. Once the offline model has been trained, it is validated using various metrics typically used for machine learning regression problems. The accuracy of the model can be assessed using Mean Absolute Error (MAE) or the Pearson correlation coefficient (R^2). The error residuals are also visualized in order to understand the overall performance of the models.

4. Results

In this section, the results obtained from the application of the framework outlined in Fig. 1 are demonstrated. The various subsections provide details of specific considerations and experiments conducted during the application of the framework to the data available in this study. Section 4.1 provides an overview of the data set used in this work. Section 4.2 describes experiments conducted to down-select the most important features for this particular problem. Section 4.3 demonstrates the advantages of using a balanced data set versus an imbalanced data set with respect to flight events for training the offline model. Section 4.4 describes the training of offline models using the knowledge from the previous three subsections. Section 4.5 presents the application of the developed framework on the second metric of interest (ground speed). Finally Section 4.6 provides a brief comparison between this paper and previous similar works in literature.

4.1. Data set

In the present work, historical de-identified FOQA data from commercial airline routine operations is utilized. The data consists of approximately 18,000 flights with over 600 parameters collected at a frequency of 1 Hz. There are six airframe groups containing two engine aircraft included in the data set (B737-group, B777-group, B757-group, MD80-group, A320-group, A330-group). The data consists of operations conducted at over 70 airports. The total size of the data set is over 600 GB with each flight averaging around 35 MB. The data from each flight can be further divided into various phases of flight such as Taxi Out, Take off, Climb, Cruise, Approach,

[4] Source: FAA Runway Safety (https://www.faa.gov/airports/runway_safety/excursion/)

[5] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

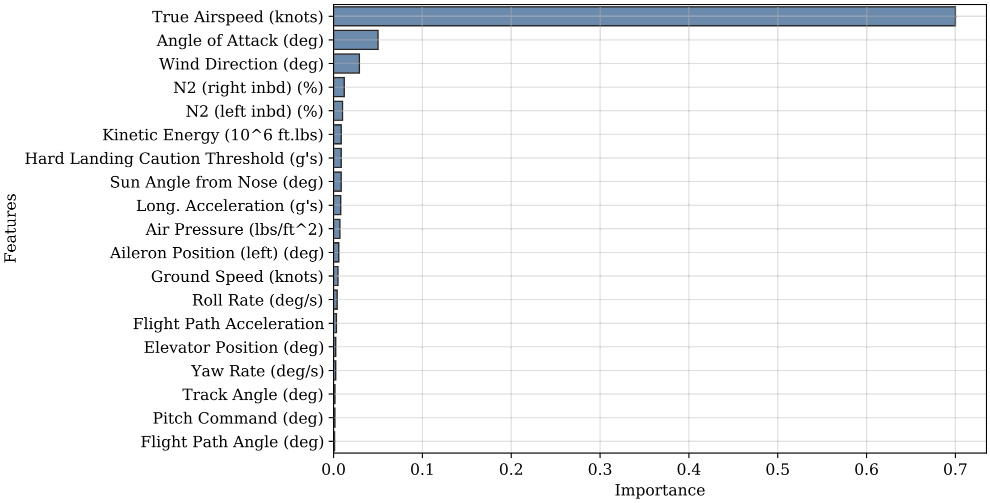


Fig. 4. Feature importance calculated by permutation importance for top 20 features of random forest model.

Landing, Roll Out, etc. Some of the parameters in the data are invariant throughout the flight (for example: the take-off and landing runway/airport information), whereas other parameters may change every second within the data (for example: airspeed, altitude, vertical speed, etc.). Due to the high dimensionality of the data and physical and analytical redundancies in sensor systems on-board an aircraft, there exists a certain amount of correlation between the parameters in the flight data. The data is extracted in the form of text files and converted to Comma Separated Variable (CSV) format along with some data cleaning (such as removal of incomplete or corrupted files etc.). Each extracted flight is thus available in the form of a large matrix in which the number of rows is equal to the total duration of the flight in seconds (between $\approx$ 3000 to 20000 rows) and the number of columns is the total number of parameters recorded ($\approx$ 623 columns).

In this paper the metric or target of interest in all cases is the **true airspeed** at touchdown (averaged over five samples) other than Section 4.5 where the metric is ground speed. This can be changed to any other metric but for the sake of demonstration in this paper, true airspeed and ground speed are used. The decision gates used for data collection are 1000 feet and 300 feet (A provides a sensitivity analysis on this choice of decision gates for the feature vectors). The typical descent rate (for the flights in the current dataset) during the approach phase is between 1000 and 1200 feet per minute (fpm) which corresponds to roughly 16–20 feet per second (fps). Therefore, collecting data every 20 feet would be approximately equivalent to collecting once per second which is the frequency of the available data. Thus data is sampled every 20 feet between the two decision gates. This approach prevents under-sampling or over-sampling which can lead to poor accuracy of models. The reason for collecting data based on feet descended rather than time is because different aircraft will take different amounts of time to descend, but the altitude increments will remain the same. Thus, during model building, the same number of samples are obtained for each flight.

4.2. Feature selection

For any predictive model building process, one of the foremost tasks is selection of appropriate features for the model (Kuhn and Johnson, 2013). As is evident from Section 4.1, over 600 parameters are available at each sample for the flight data records. However, for prediction of true airspeed and ground speed at touchdown, not all of them are equally important. Therefore, feature selection is the first step towards building an accurate prediction model. The initial set of 623 features is reduced to 300 features by removing empty parameters or metadata parameters. This step is followed by removing features that are partially empty or NaN (not-a-number) for majority of the flights during the approach phase. This reduces the number of features from 300 down to $\approx$ 97. Using these 97 parameters between the two gates during the approach phase, an initial random forest regression model is trained to predict airspeed at touchdown. The main use of this model is as a black-box for the feature selection using permutation importance method. Since this model is used for screening purposes, the model hyperparameters are frozen at the following settings:

- Minimum number of samples required to split a node = 3
- Minimum number of samples required at each leaf node = 2
- Tree Depth = 50
- Number of Estimators = 250

Fig. 4 represents the results obtained from the permutation importance algorithm. The different features are represented on the y-axis along with their normalized feature importance on the x-axis. The feature importance of a single parameter (for example airspeed) at multiple samples (for example 1000 feet, 980 feet, …, etc.) is combined into a single value to understand the overall importance at a parameter-level.

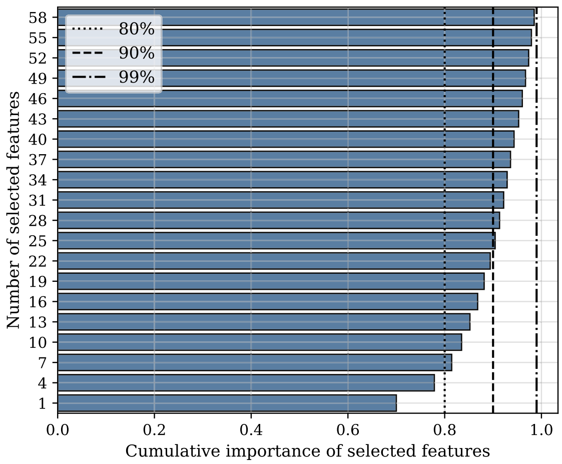


Fig. 5. Cumulative feature importance for trained random forest model.

As is evident from the figure, the most important parameter to predict true airspeed at touchdown using data between 1000 and 300 feet is the true airspeed during that time. While this is a trivial result, it validates the implementation of the framework. Similarly, the other top features identified from this exercise enable down-selecting the features further. Parameters that are expected to influence the landing true airspeed from a physical perspective are observed in the list including wind direction, ground speed, kinetic energy, etc. while some others such as landing weight are included in the larger feature set, but not in the top 20. Fig. 5 represents the cumulative feature importance up to the top *n* features – where *n* is the number of selected features. The three vertical lines represent 80, 90, and 99% cumulative feature importance. Approximately 58 features are required for capturing 99% of the feature importance and these 58 features are thus chosen to build the prediction model. It is noted that the feature importance only changes in magnitude and the cumulative feature importance does not significantly change even when the data collection gates are varied (as in Appendix A). The choice of cumulative feature importance threshold of 99% is flexible and is chosen based on a balance between complexity of the model and maximizing cumulative feature importance of chosen features. Thus, the final step of wrapper-based feature selection cuts down the number of features from 97 to 58 by retaining all features up to the 99% cumulative feature importance value.

4.3. Balancing training data

Once the features to be used in the feature vector for model building are obtained, the next experiment conducted is to ensure the model is built using a balanced data set. The data available is divided initially into two sets – the training and test set. The training set is used to train and validate the model and the test set is held back for evaluation of the model's performance. Typically, in classification problems, the concept of balanced training dataset is used to understand the number of relative samples from each class. A dataset is well-balanced if it contains approximately equal number of samples from each class. While there are no direct analogies in regression models, the data used within the training set can still influence the quality of the overall model significantly. In this section, one such major contributor for aviation safety analysis is explored and a recommendation provided for this framework.

In traditional aviation safety analysis, *events* are defined as the concurrent exceedance of one or more flight parameters beyond threshold values defined by subject-matter-experts (CAA, 2013). Events are useful to identify potential precursors to other severe states such as incidents or accidents. Typically flights with recorded in-flight events are analyzed retrospectively as they demonstrate deviation from normal behaviors. It is noted that while the events are meant to be a safety barrier, their causes can be manifold – environmental (bad weather conditions), traffic/congestion related (instructions from Air Traffic Control to follow a certain trajectory), instrument-related (sensor errors), human error, etc. Irrespective of the cause, it is understood that flights with events may behave significantly differently than those without (Li et al., 2015; Matthews et al., 2013). As such, a model aimed at predicting the touchdown metrics for safety purposes should ideally be able to predict it for flights with and without events. Therefore, in this section the model highlighted earlier is trained using two different types of training datasets – one training set with a good mix of event and non-event flights and another with randomly chosen training set. Due to the very low accident rates in commercial aviation, the amount of flights with events is usually much lower than those without events (it could be as low as 1–10% of the total number of flights (Li et al., 2015)). Therefore, when randomly selecting a training set, the likelihood of selecting many event flights is lower.

Fig. 6 shows the distribution of mean absolute error for flights from the test set contained in various event categories for the balanced (gray) and imbalanced (blue) training sets. The names of different event categories are included on the y-axis along with the total number of test flights in that event category in text next to the bars. The balanced training set is generated using the logic described earlier so as to include event flights from all categories whereas the imbalanced training set contains an average of scores obtained from random selection of training flights over 100 runs. Using a balanced training set greatly enhances the predictive power of the model for most event categories, especially for those that are relatively infrequent. The overall mean absolute error of the model calculated over the different event categories also improves from 3.43 knots to 2.50 knots. The results shown here indicate that the

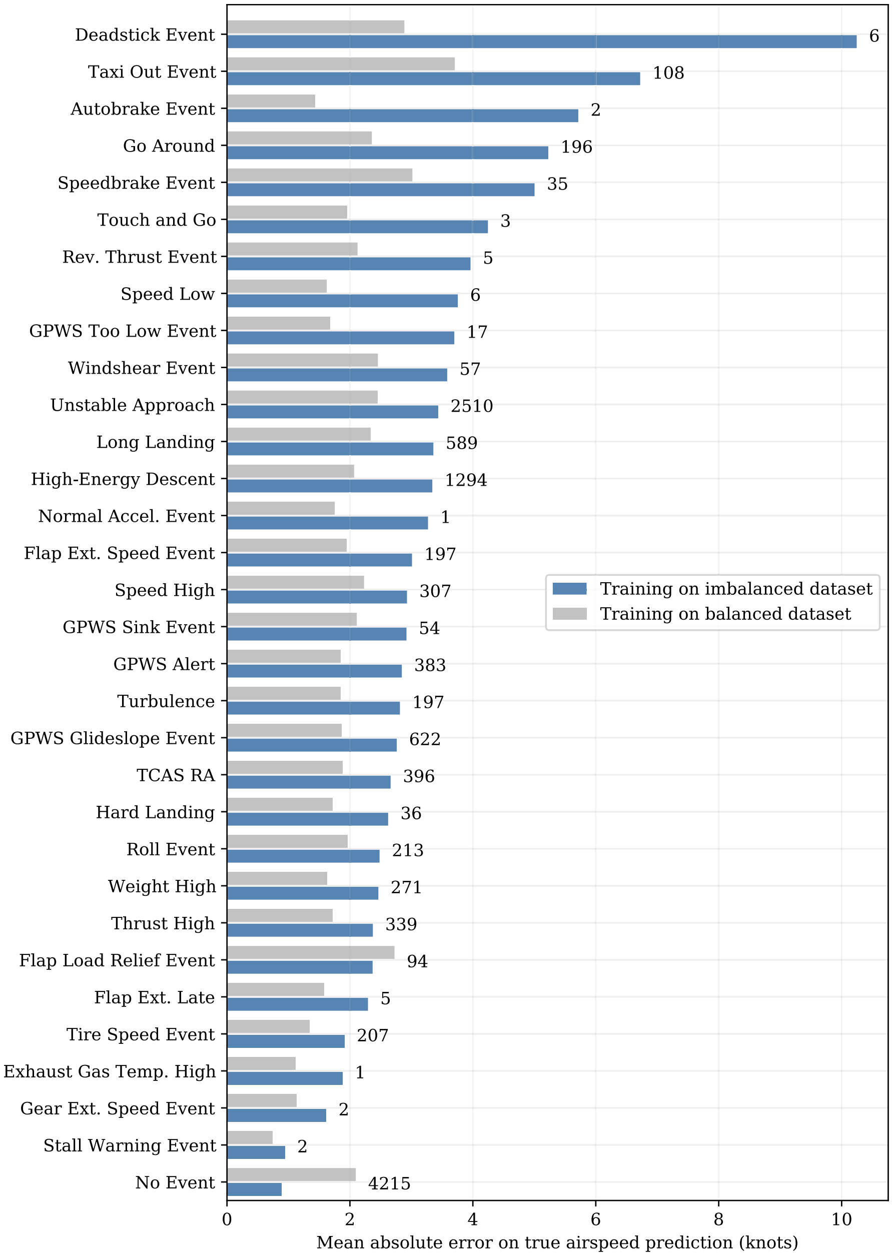


Fig. 6. Mean absolute error for true airspeed prediction using an *event*-balanced dataset versus an *event*-imbalanced dataset. The number of flights in the test dataset for the corresponding event are noted next to the error bars for each event type.

model does have a slight tendency to overfit to the non-event flights if event flights from all categories are unavailable in the training set. It is noted that the event label is only used in the training process to inform selection of training flights. The models themselves are built purely based on the raw parameters recorded in-flight. Therefore, in this work, an additional constraint is imposed while choosing the flights in the training set – at least one flight from each event category (where available) is included in the training set.

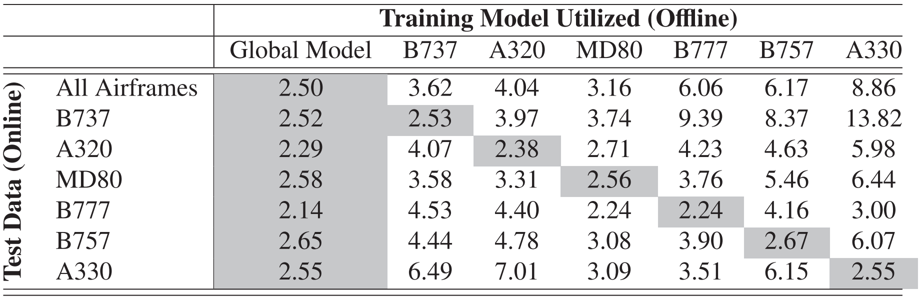
4.4. Prediction model results

Based on the findings of the previous sections, a random forest prediction model is built using the top 58 features from the approach phase sampled every 20 feet between 1000 and 300 feet using a balanced data set. The random forest model's hyperparameters are trained by varying them in a Design-of-Experiments and choosing the model with the best performance on the validation set. The hyperparameters considered and their limits are as follows:

Table 1
Table showing the total number of training flights and the proportion of flights with safety events in the data sets.

	Training Model						
	Global Model	B737	A320	MD80	B777	B757	A330
Number of training flights	800	800	800	800	330	800	562
Proportion of events in training set	95%	61%	55%	52%	51%	47%	59%

Table 2
Mean absolute error for true airspeed in knots for all training model-test data combinations.



		Training Model Utilized (Offline)						
		Global Model	B737	A320	MD80	B777	B757	A330
Test Data (Online)	All Airframes	2.50	3.62	4.04	3.16	6.06	6.17	8.86
	B737	2.52	2.53	3.97	3.74	9.39	8.37	13.82
	A320	2.29	4.07	2.38	2.71	4.23	4.63	5.98
	MD80	2.58	3.58	3.31	2.56	3.76	5.46	6.44
	B777	2.14	4.53	4.40	2.24	2.24	4.16	3.00
	B757	2.65	4.44	4.78	3.08	3.90	2.67	6.07
	A330	2.55	6.49	7.01	3.09	3.51	6.15	2.55

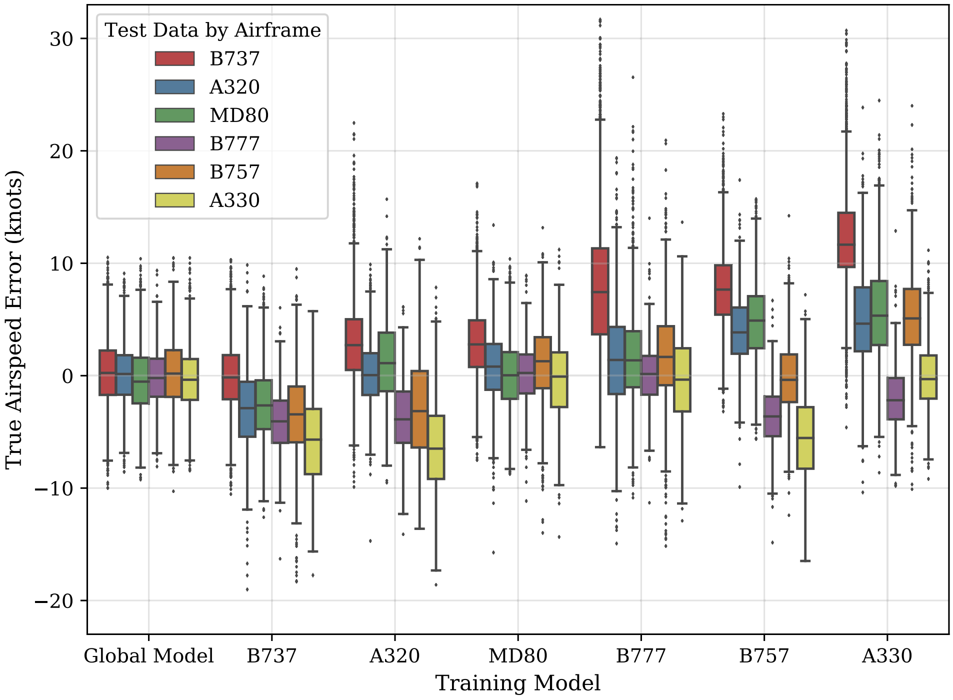
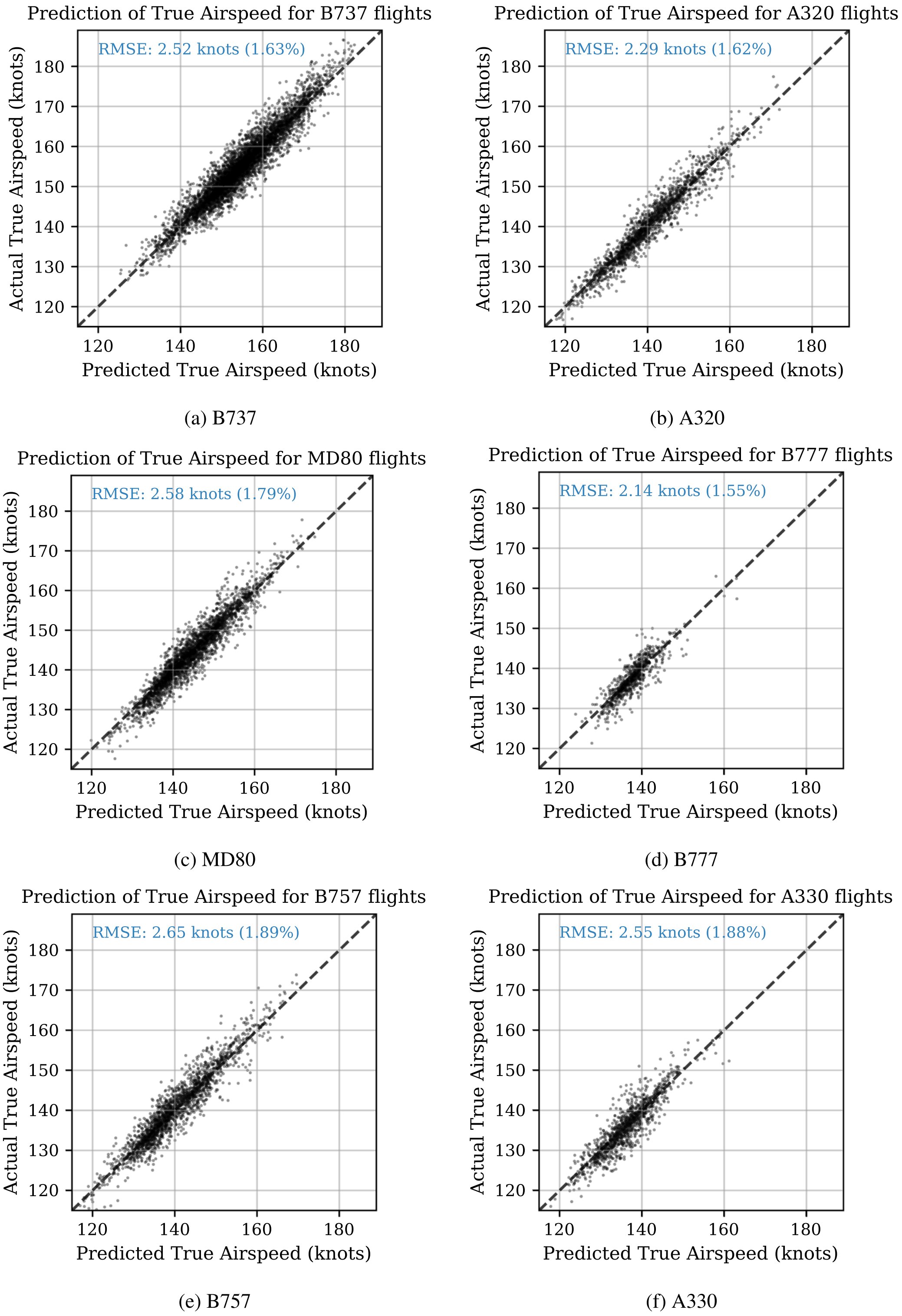


Fig. 7. Boxplots showing the true airspeed error in knots for all training model-test dataset combinations.

- Minimum number of samples required to split a node: [2, 10]
- Minimum number of samples required at each leaf node: [1, 5]
- Tree Depth: [10, 100]
- Number of Estimators: [100, 1000]

These limits are based on varying the values around the defaults available in the package while allowing for a decent exploration of the hyperparameter space. The final set of chosen hyperparameters based on predictions on the validation set is 5 minimum samples required to split a node, 3 minimum samples required at each leaf node, tree depth of 40, and 500 estimators.

The balanced training set used for the model contains approximately 6000 flights spread unequally across six different airframe



(a) B737

(b) A320

(c) MD80

(d) B777

(e) B757

(f) A330

Fig. 8. Scatterplots showing a variation of the actual versus predicted touchdown airspeed for different airframes in the test data obtained using the global prediction model.

groups. Because of this heterogeneity in airframes, a research question that needs to be addressed is whether building a separate model for each airframe is required or a single global model containing data from all airframes can accurately predict the landing airspeed. From an applicability perspective, a single global model that can account for multiple airframes is more versatile than an individual models for each airframe. If a global model works *at least as good as* any specialized fleet level model, then that precludes the training and maintenance of multiple fleet-level models. This indicates that the developed global model would be much more broadly applicable and can (in the future) absorb data from and predict for other fleets/airframes as well. Another benefit of the global model is that all airframes might not experience all types of events, but the global model can inherently apply its learning from a few airframes which do experience safety events to other airframes which may not have flights where these events were experienced.

In order to answer this question, the following approach is taken: seven different prediction models are trained, one on training data from each airframe and a global model that contains data from all available airframes. The total number of training flights and proportion of events contained in the training data is shown in Table 1. For each airframe the number of training flights is equal to the greater of eight hundred or the total number flights contained in the training dataset for that particular airframe. Eight hundred is chosen so that most models are trained with the same number of flights in order to have a fair comparison. Consequently, due to the limitations of the data available, two airframes contain fewer flights in the individual airframe-level models. The global model includes eight hundred flights from across all airframes and has majority of the events represented in the training set. Under these conditions, the models are trained and validated and are deployed on the test dataset which contains data from all airframes. The results obtained from this experiment are shown in Table 2 and plotted in Fig. 7.

Table 2 contains the results of mean absolute error (MAE) on the test dataset for various training model options. The first row contains the MAE over data from all airframes in the test set and rows 2–7 contain the MAE obtained on flights from individual airframes in the test dataset. Reading the table column-wise is an indicator of the performance of a particular model for over the different airframes in the test set whereas reading it row-wise provides an estimate of how well different models perform for test data from that particular airframe. The highlighted cells indicate the corresponding model for that test data set and its comparison to the global model. The global model performs at least as good as each fleet-level model irrespective of the test data used in all but one case (MD80). Even for this case, the MAE for the global model is only 0.78% higher than the MD80 model. Thus, it is evident from Table 2 that the global model is *sufficient* to model the landing true airspeed using the developed methodology.

Fig. 7 shows the box plots of the true airspeed prediction error (actual minus predicted) for all airframes using the different models. The central line in each box represents the median and the edges of the box indicate the quartiles of the error. The whiskers extend up to 1.5 times the inter-quartile range and the handful of points beyond the whiskers indicate outliers. The errors for the global model are generally centered at zero and evenly distributed indicating no bias in the model. The inter-quartile range is within ±10 knots for all test airframes for the global model. The global model thus performs consistently well for all test airframes whereas the individual models typically perform well for test data from the same airframe or similar-sized airframes but perform poorly on other airframes. For example, B737 model performs well on B737 test data or A320 test data but is poor at predicting touchdown speeds for A330 airframe test data. This behavior is expected as the wide-body airframes have significantly different aerodynamics than the narrow-body/single aisle airframes. These findings indicate that it is preferable to build a global model that contains data from all airframes rather than individual airframe-specific models as the global model is able to differentiate between the airframes implicitly while predicting.

The next step in the model building framework is to use all the insights from the previous experiments and build the global model for touchdown airspeed prediction. This model is built using all the training data from around 6,000 flights and the test data consists of around 10,000 flights that have not been used in any part of the model building or tuning process. The mean absolute error for this global model over the entire test dataset was reported earlier in Table 2 as **2.50 knots (1.71% relative error)**. This error is much lower than the errors reported in previous works in literature for similar prediction models (Tong et al., 2018; Diallo, 2012).

Fig. 8 contains scatter plots of actual (y-axis) versus predicted (x-axis) touchdown airspeed in knots for the test dataset for each of the different airframe groups. In each sub-plot, the global model is deployed to predict the touchdown true airspeed at the 300 feet above touchdown point and the actual touchdown speed from the full flight record is compared. The dashed line at 45 degrees in each figure represents the ideal location for all data points as that would mean no discrepancy between actual and predicted speed. The figures indicate a good overall agreement between the actual and predicted airspeed for all airframes. Most of the predictions are clustered around the diagonal dashed line. The global model performs well for both the high and low touchdown speed regions indicating that it is able to accurately predict (within 3 knots on an average) in both scenarios. All of these prediction capabilities are available at 300 feet above touchdown altitude which, according to experiments by Campbell et al. (2018), should provide sufficient time for pilots to perform a go-around maneuver if the predicted touchdown speed is deemed to be unsafe.

Figs. 7 and 8 indicate that the model is good at predicting the true airspeed for most flights with a tight inter-quartile around zero for the error. However, there are some outliers with higher residuals (between 5–10 knots absolute error). An analysis of these outliers is performed to identify the reasons for the poor predictions and potential solutions to the issue. 22 out of approximately 9000 test flights (0.24%) have an error of 10 knots or higher (which is equivalent to a relative error of ≈ 6%) using the global model. The maximum error is 12.24 knots. 20 out of the 22 outlier flights contained some event among those outlined in Fig. 6. Among the events in the 20 flights are unstabilized approach, high energy descent, turbulence, and glideslope deviations. However, it is noted that the model performance is within tight tolerance for hundreds of other flights in the test set with these events. There are no specific trends observed among the outliers related to airframe, gross weight, or other metadata characteristics. For all the outliers the median actual true airspeed is 160 knots and all outliers are within the inter-quartiles for actual true airspeed in the test set. This indicates that the outliers are not biased towards any of the extremes of the landing performance (high or low airspeed landings). These observations indicate that the outliers in this model occur because of the trade-off between the training accuracy and generalizability of the model. It is observed that among the model hyperparameters, the minimum number of samples required at each leaf node and the minimum number of samples required to split a node have an impact on the proportion of outliers. Both hyperparameters thus need to be included in the hyperparameter search for the best performing model in any future effort leveraging this methodology.

Table 3
Mean absolute error for ground speed in knots for all training model-test data combinations.

		Training Model Utilized (Offline)						
		Global Model	B737	A320	MD80	B777	B757	A330
Test Data (Online)	All Airframes	2.80	3.58	3.20	3.42	4.47	3.39	4.76
	B737	2.80	2.83	3.36	3.47	5.94	3.76	6.48
	A320	2.23	3.57	2.40	2.72	3.17	2.51	3.23
	MD80	3.03	4.18	3.55	2.97	3.98	3.66	4.18
	B777	2.29	3.33	2.47	3.45	2.42	2.43	2.91
	B757	3.05	4.52	3.22	4.15	3.40	3.13	3.55
	A330	3.04	4.56	3.28	4.03	3.36	3.34	2.98

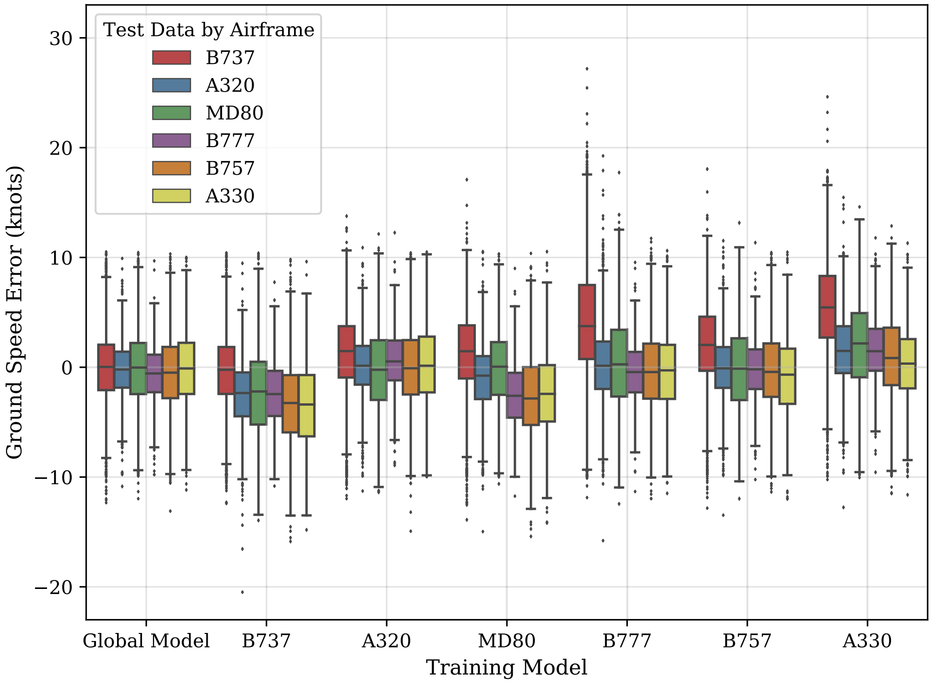


Fig. 9. Boxplots showing the ground speed error in knots for all training model-test dataset combinations.

Finally, for any predictive model that is targeted to be used in an online setting computational time during prediction is important. In this scenario, the framework can afford to incur a higher computational cost for training but needs to be extremely fast in the prediction phase. Due to the data cleaning and feature selection only ≈ 10% of the available 623 parameters are required for prediction. For the global model trained in this work, the average prediction time over the entire test set is 0.01 s on a regular desktop computer (i7 Processor, 8 GB RAM). This prediction time includes collecting and organizing data into the appropriate format, feeding it into the prediction model, and obtaining the predicted value of the landing speed. In the context of the use case for this work, this is a very favorable outcome as it provides enough time for decision-making. Due to the very fast prediction time of the developed model, further detailed studies are not conducted as it satisfies the basic requirement for online applicability. Thus, the performance of the model developed in this work is appropriate from an online prediction perspective.

4.5. Extension of model to ground speed prediction

In this section, the framework demonstrated on true airspeed is applied to ground speed, another landing performance metric of importance from the perspective of runway excursion risk. Using a similar process to that outlined for true airspeed, top features retaining 99% of cumulative feature importance are identified for the ground speed prediction model. The hyperparameters are tuned for the models and the same balanced training and test data sets as the true airspeed model are used to maintain consistency between the two models. The results obtained for this model are highlighted in Table 3. The results indicate a similar trend to that observed for true airspeed with the difference being slightly higher mean absolute error for the ground speed model (2.80 knots) compared to the true airspeed model (2.50 knots). The global ground speed model also performs as good as individual models (except MD80, A330

Table 4
Table showing the comparison of the present approach with similar work in literature.

Measure	This Paper	Tong et al. (2018)	Diallo (2012)
True Airspeed (RMSE)	2.62 knots (1.79%)	N/A	3.5 knots
Ground Speed (RMSE)	2.98 knots (2.10%)	3.5 knots	N/A
Advance Prediction	300 feet (≈ 18 s)	1 s	Retrospective Model
Valid for Flights with Events	Yes (≈ 35 events)	N/A	Yes (1 Event: Gusts)
Number of Airframes	6	N/A	1
Number of Airports	70	N/A	1

models where it is ≈ 2% worse) and is thus sufficient to predict ground speed at landing.

Fig. 9 shows the boxplot of the ground speed error for the various training models as applied to the test data set. A similar trend is observed to that of the true airspeed where the inter-quartile range is within ± 10 knots for the global model. Thus, from Table 3 and Fig. 9 it is evident that the developed method yields a global model that is significantly better at prediction of ground speed than previous work in literature and provides the prediction at an earlier point during the approach phase.

4.6. *Comparison to previous work*

A summary of the comparison of the present approach with similar works identified in literature is presented in Table 4. It is evident from the table that this paper significantly improves upon the state of the art in terms of root mean square error for two types of landing metrics, advance prediction (how far ahead of the touchdown is the prediction available), variety of flights (flights with safety events and no events), and robustness (applicability of a single model to multiple airframes and airports). Additionally, this work presents a thorough sensitivity analysis and detailed descriptions of each step of the methodology which are lacking in some previous work. The most significant of these is the ability to predict the true airspeed and ground speed at 300 feet above touchdown to within a few knots. This is because predicting the true airspeed and ground speed accurately using the data from the entire approach phase up to the touchdown point does not provide any decision-making capability and uses much more data than that used in the present work.

5. Conclusion

In this paper, a novel framework for the analysis of aviation flight data using supervised machine learning is demonstrated. The elements of the framework are described in detail along with a practical use-case for the online prediction of aircraft landing true airspeed and ground speed during the approach phase. Historical data from commercial airline operations is used to train a global prediction model using random forest regression offline. The trained model is then deployed on new test data to test the capability of prediction in an online setting.

A number of innovations and improvements over existing methods are provided in this work. During the development of the framework, a method of automated feature selection for the prediction model is demonstrated with permutation importance measure. This enables reducing the total number of parameters used in the process significantly and decreases the dependency on a large parameter set for prediction of the metric. The effect of balancing the training dataset with respect to flights containing safety events is also demonstrated to further improve the accuracy of the model. A sensitivity analysis is conducted on appropriate windows of data to be used for prediction of the landing metrics which indicated that the 300 foot above touchdown gate was the best from accuracy perspective while providing enough time for pilots to make decisions. The developed model has a good computational performance for potential online application by providing the predictions within a second for all test flights considered. Finally, the framework is flexible in that it can be adapted for prediction of other metrics of interest as demonstrated by the adaptation for prediction of ground speed. One of the limitations of the developed framework is the outliers identified in the results section for flights with absolute value of prediction error slightly over 10 knots. The outliers are a result of the trade-off between training accuracy and generalizability of the model. Thus, while the proportion of outliers is small (0.24%), the predictions in an online setting should be paired with other decision-making to make a well-informed, safe go-around or landing decision.

One of the main benefits benefit of this framework is that while stabilized approach criteria can become operationally complex to understand and implement, a direct prediction of landing airspeed or ground speed from this framework can provide more actionable information to the pilots, operators, or air traffic control to improve their decision-making capability. In future work, the capability of this framework will be further extended to include consensus predictions for scenarios such as sensor failures when a single model might not provide accurate predictions. Additionally, the demonstration of the framework for other metrics of interest will also be presented. Investigation also needs to be performed on the mechanisms for enabling online deployment of such a developed model in-flight.

CRediT authorship contribution statement

Tejas G. Puranik: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing - original draft, Writing - review & editing. **Nicolas Rodriguez:** Data curation, Software, Visualization, Writing - original draft. **Dimitri N. Mavris:** Supervision, Project administration, Resources, Writing - review & editing.

Appendix A. Sensitivity analysis for decision gate selection

In this section, a sensitivity analysis of decision gate selection is performed. Fig. 10 shows the different combinations for the two decision gates that are available. The prediction model is trained for data collected between each combination of decision gates using the trimmed feature set and event-balanced training set and the mean absolute error is obtained for the test data in each case. The heatmap represents the MAE for each combination of gates. As is evident from the heatmap, for all the combinations, the errors are usually low, with the location of the second gate (where the prediction is actually made) having a higher impact on accuracy. Interestingly, using data from further behind in the approach while keeping the second gate constant does not improve predictive capabilities – for example, error is higher for [3000, 300] feet combination than [1000, 300] feet combination. Based on the results from the figure and on the understanding of decision making during approach phase, the decision gate combination of [1000, 300] feet is chosen for this work because it affords the advantage of lowest error while still being within acceptable reaction time for go-around decision making according to Campbell et al. (2018).

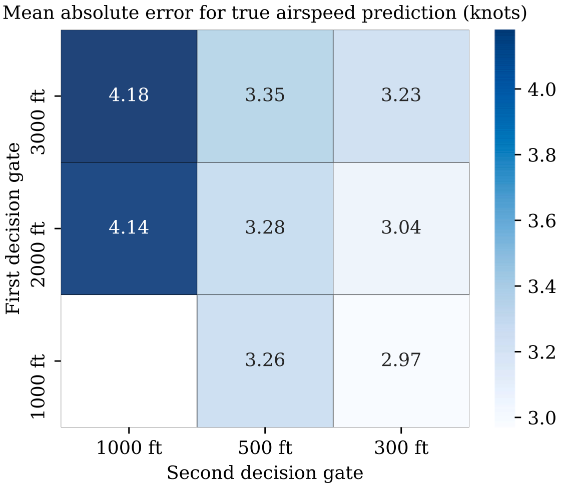


Fig. 10. Heatmap showing the mean absolute error on the test dataset for different combinations of decision gate choices.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.trc.2020.102819.

References

Advisory Circular, 2004. 120–82 – Flight Operational Quality Assurance. https://www.faa.gov/regulations_policies/advisory_circulars/index.cfm/go/document.information/documentID/23227.

Amidan, B.G., Ferryman, T.A., 2000. APMS SVD Methodology and Implementation, Technical Report, U.S. Department of Energy PNWD-3026. doi:10.2172/753847.

Basora, L., Olive, X., Dubot, T., 2019. Recent advances in anomaly detection methods applied to aviation. Aerospace 6 (11), 117.

Belcastro, C.M., Jacobson, S.R., 2010. Future integrated systems concept for preventing aircraft loss-of-control accidents. In: AIAA Guidance, Navigation, and Control Conference. Paper No. AIAA-2010-8142.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Budalakoti, S., Srivastava, A.N., Otey, M.E., 2009. Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety. IEEE Trans. Syst. Man Cybernet. Part C: Appl. Rev. 39, 101–113. https://doi.org/10.1109/TSMCC.2008.2007248.

CAA, 2013. Civil Aviation Authority – Flight Data Monitoring CAP 739 Second Edition, 2013. ISBN 978-0-11792-840-4. https://publicapps.caa.co.uk/docs/33/CAP739.pdf.

Campbell, N., 2020. Flight data analysis – an airline perspective. In: Australian and New Zealand Societies of Air Safety Investigators Conference.

Campbell, A., Zaal, P., Schroeder, J.A., Shah, S., 2018. Development of possible go-around criteria for transport aircraft. 2018 Aviation Technology, Integration, and Operations Conference. Paper Number: AIAA-2018-3198.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: a survey. ACM Comput. Surv. (CSUR) 41. https://doi.org/10.1145/1541880.1541882.

Das, S., Matthews, B.L., Srivastava, A.N., Oza, N.C., 2010. Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 47–56. https://doi.org/10.1145/1835804.1835813.

Diallo, O.N., 2012. A predictive aircraft landing speed model using neural network. In: 2012 IEEE/AIAA 31st Digital Avionics Systems Conference (DASC), IEEE, pp. 3D2–1.

Federal Aviaition Administration Aerospace Forecasts Fiscal Years 2016–2036, 2017. https://www.faa.gov/data_research/aviation/aerospace_forecasts/media/FY2016-36_FAA_Aerospace_Forecast.pdf. Retrieved: 10/2019.

Federal Aviation Administration, 2011.14 CFR §121.344 Digital Flight Data Recorders for Transport Category Airplanes, https://www.ecfr.gov/cgi-bin/text-idx?SID=b42b5be68aa3c7da5b85e2c60277e054&mc=true&node=se14.3.121_1344&rgn=div8.

Federal Aviation Administration Advisory Circular 120–71a, 2003. Advisory Circular. Retrieved: 10/2019.

Federal Aviation Administration - Aviation Safety Information Analysis and Sharing (ASIAS), 2017. http://www.asias.faa.gov. Retrieved: 10/2019.

Flight Safety Foundation, 2000. ALAR Briefing Note 7.1 – Stabilized Approach, Technical Report. Retrieved: 10/2019.

Flight Safety Foundation, 2000. ALAR Briefing Note 4.2 – Energy Management, Technical Report. Retrieved: 10/2019.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Hegde, J., Rokseth, B., 2020. Applications of machine learning methods for engineering risk assessment–a review. Saf. Sci. 122, 104492.

Iverson, D.L., 2004. Inductive System Health Monitoring, Technical Report, National Aeronautics and Space Administration. https://ntrs.nasa.gov/search.jsp?R=20040068062.

Jarry, G., Delahaye, D., Nicol, F., Feron, E., 2020. Aircraft atypical approach detection using functional principal component analysis. J. Air Transp. Manage. 84, 101787.

Jarry, G., Delahaye, D., Feron, E., 2020. Approach and landing aircraft on-board parameters estimation with lstm networks. In: 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT), IEEE, pp. 1–6.

Kuhn, M., Johnson, K., 2013, An introduction to feature selection. In: Applied Predictive Modeling, Springer, pp. 487–519.

Lee, S., Park, W., Jung, S., 2014. Fault detection of aircraft system with random forest algorithm and similarity measure. Sci. World J. 2014.

Lee, H., Madar, S., Sairam, S., Puranik, T.G., Payan, A.P., Kirby, M., Pinon, O.J., Mavris, D.N., 2020. Critical parameter identification for safety events in commercial aviation using machine learning. Aerospace 7, 73.

Li, G., 2010. Machine learning in fuel consumption prediction of aircraft. In: 9th IEEE International Conference on Cognitive Informatics (ICCI'10), IEEE, pp. 358–363.

Li, L., Das, S., John Hansman, R., Palacios, R., Srivastava, A.N., 2015. Analysis of flight data using clustering techniques for detecting abnormal operations. J. Aerosp. Inf. Syst. 12, 587–598.

Li, L., Hansman, R.J., Palacios, R., Welsch, R., 2016. Anomaly detection via a gaussian mixture model for flight operation and safety monitoring. Transp. Res. Part C: Emerg. Technol. 64, 45–57.

Logan, T.J., 2008. Error prevention as developed in airlines. Int. J. Radiat. Oncol. Biol. Phys. 71, S178–S181.

Lujan-Moreno, G.A., Howard, P.R., Rojas, O.G., Montgomery, D.C., 2018. Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. Expert Syst. Appl. 109, 195–205.

Martınez, D., Fernández, A., Hernández, P., Cristóbal, S., Schwaiger, F., Nunez, J.M., Ruiz, J.M., 2019. Forecasting unstable approaches with boosting frameworks and lstm networks. In: 9th SESAR Innovation Days.

Matthews, B., Das, S., Bhaduri, K., Das, K., Martin, R., Oza, N., 2013. Discovering anomalous aviation safety events using scalable data mining algorithms. J. Aerosp. Inf. Syst. 10, 467–475.

Moriarty, D., Jarvis, S., 2014. A systems perspective on the unstable approach in commercial aviation. Reliab. Eng. Syst. Saf. 131, 197–202.

Mugtussids, I.B., 2000. Flight Data Processing Techniques to Identify Unusual Events. Ph.D. thesis. Virginia Polytechnic Institute and State University. http://hdl.handle.net/10919/28095.

Petukhova, T., Ojkic, D., McEwen, B., Deardon, R., Poljak, Z., 2018. Assessment of autoregressive integrated moving average (arima), generalized linear autoregressive moving average (glarma), and random forest (rf) time series regression models for predicting influenza a virus frequency in swine in Ontario, Canada. PloS One 13, e0198313.

Puranik, T., Mavris, D., 2018. Anomaly detection in general-aviation operations using energy metrics and flight-data records. J. Aerosp. Inf. Syst. 15, 22–35.

Puranik, T.G., Mavris, D.N., 2019. Identification of instantaneous anomalies in general aviation operations using energy Metrics. J. Aerosp. Inf. Syst. 17, 51–65.

Puranik, T., Harrison, E., Min, S., Jimenez, H., Mavris, D., 2016. General aviation approach and landing analysis using flight data records. In: 16th AIAA Aviation Technology, Integration, and Operations Conference. Paper No. AIAA 2016–3913, doi:10.2514/6.2016-3913.

Puranik, T., Jimenez, H., Mavris, D., 2017. Energy-based metrics for safety analysis of general aviation operations. J. Aircraft 54, 2285–2297.

Rao, A.H., Marais, K., 2015. Identifying high-risk occurrence chains in helicopter operations from accident data. In: 15th AIAA Aviation Technology, Integration, and Operations Conference. Paper No. AIAA 2015–2848, doi:10.2514/6.2015-2848.

Rao,A.H., Puranik, T.G., 2018. Retrospective analysis of approach stability in general aviation operations. In: 18th AIAA Aviation, Technology, Integration, and Operations Conference, Atlanta, GA. June.

Reason, J., 2000. Safety paradoxes and safety culture. Injury Control Saf. Promotion 7, 3–14. https://doi.org/10.1076/1566-0974(200003)7:1;1-V;FT003.

Schuet, S., Lombaerts, T., Acosta, D., Kaneshige, J., Wheeler, K., Shish, K., 2017. Autonomous flight envelope estimation for loss-of-control prevention. J. Guid. Control Dyn. 40, 847–862.

Sherry, L., Wang, Z., Kourdali, H.K., Shortle, J., 2013. Big data analysis of irregular operations: aborted approaches and their underlying factors. In: 2013 Integrated Communications, Navigation and Surveillance Conference (ICNS), IEEE, pp. 1–10.

Statistical Summary of Commercial Jet Airplane Accidents - Boeing Commercial Airplanes, 2017. http://www.boeing.com/resources/boeingdotcom/company/about_bca/pdf/statsum.pdf. Retrieved: 10/2019.

Subramanian, S.V., Rao, A.H., 2018. Deep-learning based time series forecasting of go-around incidents in the national airspace system. In: 2018 AIAA Modeling and Simulation Technologies Conference. Paper Number: AIAA 2018-0424.

Tong, C., Yin, X., Wang, S., Zheng, Z., 2018. A Novel Deep learning method for aircraft landing speed prediction based on cloud-based sensor data. Fut. Gen. Comput. Syst. 88, 552–558.

Tong, C., Yin, X., Li, J., Zhu, T., Lv, R., Sun, L., Rodrigues, J.J., 2018. An innovative deep architecture for aircraft hard landing prediction based on time-series sensor data. Appl. Soft Comput. 73, 344–349.

Unpingco, J., 2016. Python for Probability, Statistics, and Machine Learning, first ed. Springer.

Wang, Z., Sherry, L., Shortle, J., 2015. Airspace risk management using surveillance track data: stabilized approaches. In: 2015 Integrated Communication, Navigation and Surveillance Conference (ICNS), pp. W3–1–W3–14.

Wang, Z., Sherry, L., Shortle, J., 2016a. Improving the nowcast of unstable approaches. In: 8th International Conference on Research in Air Transportation.

Wang, Z., Sherry, L., Shortle, J., 2016b. Feasibility of using historical flight track data to nowcast unstable approaches. In: 2016 Integrated Communications Navigation and Surveillance (ICNS), IEEE, pp. 4C1–1.

Zhang, H., Zhu, T., 2018. Aircraft hard landing prediction using lstm neural network. In: Proceedings of the 2nd International Symposium on Computer Science and Intelligent Control, ACM, p. 28.