DOT/FAA/TC-23/43

Federal Aviation Administration
William J. Hughes Technical Center
Aviation Research Division
Atlantic City International Airport
New Jersey 08405

Machine Learning Approach for the Evaluation of Degraded Wheel Braking Performance on Contaminated Runways

August 2023

Final report

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The U.S. Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the funding agency. This document does not constitute FAA policy. Consult the FAA sponsoring organization listed on the Technical Documentation page as to its use.

This report is available at the Federal Aviation Administration William J. Hughes Technical Center's Full-Text Technical Reports page: actlibrary.tc.faa.gov in Adobe Acrobat portable document format (PDF).

Technical Report Documentation Page

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/TC-23/43		

4. Title and Subtitle	5. Report Date
Machine Learning Approach for the Evaluation of Degraded Wheel Braking Performance on Contaminated Runways	August 2023
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
HyunKi Lee, Wenxin Zhang, Tejas Puranik, David Anvid, Alexia Payan, Michelle Kirby, Dimitri Mavris, Carter Tegen	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Georgia Institute of Technology, Atlanta, GA	
	11. Contract or Grant No.

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Aviation Research Division – ANG-E2 William J. Hughes Technical Center Federal Aviation Administration Atlantic City International Airport, NJ 08405	Final Report
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

One of the most challenging issues in the airline community has been the safe operations of aircraft during the landing phase of flight. Several entities agree that runway excursion is the most frequent type of landing incident or accident, and studies on this topic have shown that a significant contributor is the condition of the runway surface at the time of landing, as well as the potential for degraded braking performance.

In this report, several techniques leveraging machine learning and real-world data are explored to gain insights into critical parameters and conditions of potentially elevated levels of risk during a degraded braking event. Various sources of data are collected and processed. A data fusion pipeline is developed to integrate the various heterogenous sets of data investigated for subsequent studies. Several types of data analyses are conducted in order to gain insights into the different runway and weather factors that might affect aircraft braking performance at landing. A statistical model is developed based on Flight Operations Quality Assurance (FOQA) datasets to study the spool-down time of thrust reversers to demonstrate how a given set of data may be inferred when a direct measurement of aircraft behavior is not available. Then, various machine learning techniques are employed to further study the dataset of interest.

To conclude, a "gap analysis" roadmap is created to identify critical future research areas related to aircraft degraded braking on contaminated runways as part of the wider topic of runway safety. It features four main areas of interest (airport, aircraft, pilot, and weather) with a focus on the "Airport" area, and explores potential avenues for data collection, instrumentation, testing and validation, and data analysis/machine learning techniques application and development.

17. Key Words	18. Distribution Statement
	This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified		

Acknowledgements

The authors would like to thank Somil Shah, Angela Campbell, Hossein Eghbali, Paul Giesman, Raymond Zee, Michel Hovan, Charles “Cliff” Johnson, and Troy Brown for their feedback and comments.

Contents

Figures

Tables

Acronyms

Acronym	Definition
ABAR	Aircraft Braking Action Report
AC	Advisory Circular
AIREP	Aircraft Reports
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
ARC	Aviation Rulemaking Committee
ASOS	Automated Surface Observing Systems
ASTM	American Society for Testing and Materials
ATC	Air Traffic Control
ATIS	Automatic Terminal Information Service
CCLP	Compact Clustering via Label Propagation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DT	Decision Tree
DT-MIL	Deep Temporal Multiple Instance Learning
EASA	European Aviation Safety Agency
EPS	Epsilon
ESDU	Engineering Sciences Data Unit
FAA	Federal Aviation Administration
FAR	Federal Aviation Regulations
FICON	Field Condition Report
FOQA	Flight Operations Quality Assurance
GBDT	Gradient Boosting Decision Tree
GRU	Gated Recurrent Unit
GRVD	Grooved
ICAO	International Civil Aviation Organization
IGGA	International Grooving and Grinding Association
LSTM	Long Short-Term Memory
LiR	Linear Regression
LoR	Logistic Regression
MKAD	Multiple Kernel Anomaly Detection
ML	Machine Learning

MLP	Multi-layer Perceptron
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmosphere Administration
NOTAM	Notice to Airmen
NTSB	National Transportation Safety Board
OPTICS	Ordering Points to Identify the Clustering Structure
PFC	Porous Friction Course
PIREPs	Pilot Reports
PIREP BA	Pilot Reported Braking Action
PTYPE	Precipitation Type
QAR	Quick Access Recorder
RCAM	Runway Condition Assessment Matrix
RF	Random Forest
RNN	Recurrent Neural Networks
RTO	Rejected Take-off
RMSE	Root Mean Square Error
RWYCC	Runway Condition Code
SUM_SQ	Sum of Squares Error
SVM	Support Vector Machine
TALPA	Takeoff and Landing Performance Assessment
T-SNE	T-distributed Stochastic Neighbor Embedding
UTC	Universal Time Coordinated
XGBOOST	Extreme Gradient Boosting

Executive summary

Maintaining safe operations during the landing phase of an aircraft flight has consistently ranked as one of the most challenging issues for airlines, pilots, and regulators. According to the Flight Safety Foundation, a runway excursion is the most frequent type of landing incident or accident. Studies regarding runway excursions have shown that a significant contributor is the runway surface condition at the time of landing and the potential for degraded braking performance. Runway conditions such as wet, standing water or other contaminants are the most variable and difficult to quantify, in particular when it comes to the creation of predictive modeling tools of aircraft performance at the time of arrival.

In this report, techniques leveraging machine learning and real-world data are explored to gain insights into critical parameters and conditions of potentially elevated levels of risk during a degraded braking event. Several distinct, yet related, topics regarding landing safety are investigated and presented.

First, the collection and processing of various sources of data are presented, and a data fusion pipeline that integrates heterogeneous data sources is built as a foundation for subsequent studies.

Second, data analyses are conducted with respect to subsets of the fused dataset, and insights are drawn about how different runway/weather factors might affect braking performance.

Third, a statistical model is established based on Flight Operations Quality Assurance (FOQA) datasets to simulate the spool-down time of thrust reversers. This provides a demonstration of how select data could be inferred when a direct measurement of aircraft behavior is not available.

Then, various machine learning techniques are applied to the available datasets to study the degraded aircraft wheel braking performance on contaminated runways. Clustering, an unsupervised machine learning technique, is used first to group landing flights and explore various flight metrics related to the aircraft performance during approach and landing. Decision trees, a supervised learning technique, is used next to interpret the results of the clustering process, and analyses regarding outliers in the landing flights are conducted. Other supervised learning techniques, such as random forest and extreme gradient boosting, are subsequently used as classification methods to quantitatively assess the pilot braking actions and runway conditions. Artificial neural networks are also built to evaluate aircraft acceleration during landing based on relevant flight data.

Finally, a “gap analysis” roadmap is developed to identify critical paths of future avenues for research on aircraft degraded braking on contaminated runways as part of the wider topic of runway safety. It includes four main areas (airport, aircraft, pilot, and weather) with a focus on the “Airport” area, and explores topics such as data collection, instrumentation, testing and validation, and data analysis/machine learning techniques application and development.

1 Introduction

1.1 Background and motivation

Aircraft landing safety is a major focus in the aviation industry. According to information from the Federal Aviation Administration (FAA) and the National Transportation Safety Board (NTSB), runway overruns during the landing phase of flight account for approximately ten incidents or accidents every year with varying degrees of severity, with many accidents resulting in fatalities[1]. Among the common conditions identified are the dynamics of a tailwind approach; improper pilot technique for the conditions; or landing on non-dry runways that lead to increased stopping distances during the landing phase. It is therefore important to study the effectiveness of aircraft braking on contaminated runways to quantify increases in the landing field length required to successfully stop an aircraft on the available runway.

On December 8th, 2005, Southwest Airlines Flight 1248 overran runway 13C at Chicago Midway Airport (MDW) after landing on a runway contaminated with snow and slush. The Boeing 737-700 aircraft exited the end of the runway and went through an airport perimeter fence, striking an automobile and resulting in a fatality. This accident brought into focus the disparities on operational procedures and the need for improvements to existing practices. The FAA launched a review of existing procedures with the collaboration of airports, operators, and aircraft manufacturers. The resulting Takeoff and Landing Performance Assessment (TALPA) Aviation Rulemaking Committee (ARC) produced significant changes to the way aircraft braking is evaluated and operationally addressed.

Bolstered by the availability of on-board sensors and data collection programs, large amounts of data are being collected from routine aviation operations. There exists an opportunity to leverage large volumes of routine flight data in order to enhance the understanding of aircraft performance on dry and non-dry runways to further improve safety. This research effort addresses the application of big data analytics/machine learning techniques to address the problem of aircraft wheel braking performance on contaminated runways.

The purpose of this research is to develop new analysis tools and techniques leveraging machine learning and real-world data to obtain insights into critical parameters and conditions that proactively point towards potentially elevated levels of risk of degraded braking performance. The main idea is to use unsupervised machine learning techniques to identify and classify

[1] Federal Aviation Administration, “Runway Overrun Prevention,” Advisory Circular 91-79A, Nov. 2007

degraded braking scenarios versus normal conditions. Following this classification, supervised learning methods can be used to potentially predict when an aircraft has an increased risk of experiencing degraded braking prior to touchdown on a contaminated runway. The machine learning approach is expected to complement full-scale testing and other efforts to identify reduced aircraft wheel braking effectiveness seen on contaminated runways.

1.2 Research objectives and benefits

Considering the aforementioned motivating factors, the overall objective of this research is to:

- *Develop new analysis tools and techniques leveraging machine learning and real-world data to gain insights into conditions that proactively point towards potentially elevated levels of risk of degraded aircraft braking on contaminated runways.*

The main elements of this research are thus identified as follows:

- Use of unsupervised machine learning techniques to identify degraded braking scenarios versus normal conditions.
- Use of supervised machine learning methods to potentially predict when an aircraft has an increased risk of experiencing degraded braking on a contaminated runway.
- Identification of critical parameters for degraded aircraft braking scenario.

The main benefits expected from this research effort are the derivation of metrics from data and models to provide insights into the operational aspects of aircraft braking performance; a data-driven model to categorize and potentially predict risk factors and better characterize increased landing field length requirements on non-dry runways; and comparisons between aircraft braking performance in controlled flight tests versus routine operations.

1.3 Major areas of research

Six major areas of research depicted in Figure 1 have been identified and conducted to satisfy the research objectives identified earlier.

Figure 1. Major areas of research

In the subsequent sections of the report, each major area of research is elaborated further along with the corresponding results and findings.

2 Review of research and literature

The investigation of aircraft braking on non-dry surfaces goes back to the early 1960s, when actual aircraft testing was conducted by the National Aeronautics and Space Administration (NASA) using the Convair 880 as a testbed. The data gathered was a component of the work done in this area for the next 40 years. The seminal work by Mr. Thomas Yager of NASA Langley Research Center has resulted in over 50 articles and papers (starting in 1968) about runway friction, tire wear, and wheel braking effectiveness. Much of this early work in addition to data internally developed by Boeing, McDonnell Douglas, and Airbus became the basis for the Takeoff and Landing Performance Assessment (TALPA) operational procedure guidance. It should be noted that all of these materials were physics-based, Newtonian mechanics models supplemented with limited parametric testing.

This research is meant to extend these previous works by applying state-of-the-art data fusion, data analytics, and machine learning techniques to the large amounts of data made available by the collection of routine aviation operations, as well as various other sources of publicly available data.

The following sections summarize the literature review performed to obtain a solid understanding of existing runway safety related research, including braking characteristics and physics modeling, event prediction and metrics modeling, runway safety characteristics, as well as machine learning methods, and previous work performed at the Aerospace Systems Design Laboratory (ASDL) at the Georgia Institute of Technology that could be leveraged for this project. This review aims to provide sufficient knowledge about runway safety from both the physics and data science perspectives.

2.1 Physics of braking and braking characteristics

The Engineering Sciences Data Unit (ESDU) in the United Kingdom developed tire-force prediction methods based on the physics and chemistry of visco-elastic materials (rubber) and structural and fluid dynamics theory, which applied to single and multi-wheel undercarriages and correlated well with data from aircraft testing (ESDU, 2017).

In the report (ESDU, 2017), the ESDU discusses detailed formulas and figures for estimating braking force and summarizes various types of hydroplaning, including dynamic, viscous, and rubber reversion hydroplaning. This provides the basis for the U.S. Department of Transportation to promote a standardized way of analyzing takeoff performance and a standardized vocabulary for concepts related to takeoff performance in AC_25-31 (FAA, 2015). Braking coefficient estimates for different surface conditions, and presence of loose contaminants was also provided in this Advisory Circular, as shown in Figure 2 (FAA, 2015).

Runway Surface Condition Description	Wheel Braking Coefficient
• Frost • Wet (includes damp and ⅛″ (3 mm) depth or less of water) ⅛″ (3 mm) depth or less of: • Slush • Dry snow • Wet snow	Per method defined in § 25.109(c).
-15 °C and colder outside air temperature: • Compacted snow	0.20[1]
• Wet ("slippery when wet" runway) • Dry snow or wet snow (any depth) over compacted snow Greater than ⅛″ (3 mm) depth of: • Dry snow • Wet snow Warmer than -15 °C outside air temperature: • Compacted snow	0.16[1]
Greater than ⅛″ (3 mm) depth of: • Water • Slush	(1) For speeds below 85% of the hydroplaning speed[2]: 50% of the wheel braking coefficient determined in accordance with § 25.109(c), but no greater than 0.16; and (2) For speeds at 85% of the hydroplaning speed[2] and above: 0.05[1].
• Ice	0.08[1]

Figure 2. Wheel braking coefficients as a function of runway surface condition

The International Grooving and Grinding Association (IGGA) (IGGA, 2009) also reported experiments with grooved runways and concluded that runway groove reduces all types of skids on wet or flooded runways, and prevents drift at touchdown in flooded areas due to high cornering forces. The European Aviation Safety Agency (EASA) further showcased a lack of standards and requirements on operational friction and runway condition reporting at airports, reviewed a variety of issues on in-person runway condition reporting, offered solutions for airports to implement, and conducted comprehensive analyses on current practices for assessing runway performance (EASA, 2010). EASA also discussed operational friction under different temperatures and precipitation types.

Yager (Yager, 2013) described several promising means for improving tire/runway water drainage capability, brake system efficiency, and pilot training to help optimize aircraft traction performance on wet runways. Gerard Van Es (Es G. V., 2017) obtained and analyzed flight test data on braking performance on water contaminated runways to derive effective braking friction for different ground speeds and establish contamination drag levels. This provided several insights into the hydroplaning characteristics under un-braked and braked conditions. The same

author also investigated the modeling of dynamic hydroplaning speed, and showed that it improves with surface texture.

Several prediction models for various tires were also generated (Es G. V., 2001). Daidzic (Daidzic, 2017) developed a theoretical model of the maximum braking energy and the related VMBE (Maximum Brake Energy speed) speed for T-category airplanes and showed the effects of elevated temperatures on brake usage, and consequently braking time and distance. Pasindu (2011) developed a simulation model to calculate braking distance on wet pavement based on engineering mechanics and fluid dynamics theory. The simulation results illustrated the calibration and validation of the tire model, followed by computation of braking distances under different operating conditions of wheel load, tire inflation pressure, landing speed, and water-film thickness. Wahi (Wahi, 2012) developed a simulation model to represent the airplane dynamics under braking action, where the basic equations of motion for a rigid-body airplane with the forward, vertical, and pitch degrees of freedom were used.

Roginski (2012) took a manufacturer's perspective to discuss runway friction and aircraft performance. Definitions such as microtexture providing frictional properties for aircraft operating at low speeds and macrotexture providing frictional properties for aircraft operating at high speeds were described. He also explained that deviations in speed, wind, touchdown point, and delayed use of deceleration devices may become an issue in combination with worse than expected braking.

2.2 Event prediction and metrics modeling

Lv (2018) used several machine learning models to analyze 6,395 flight data records between March to December in 2016 and understand the risk of runway overrun. An "overrun dangerous line" was constructed to evaluate the overrun risk, which was then used to classify flights as "risky" and "normal", and determine in what regions of the runway the risk of the overrun was the largest. Machine learning techniques such as Support Vector Machine (SVM), logistic regression, and random forest were used to build the statistical models. Lv concluded that random forest was the most suitable model, and that the approach phase and the brake application (or thrust reversers application) phase are associated with the largest overrun risk after touchdown or before taxiing on the runway.

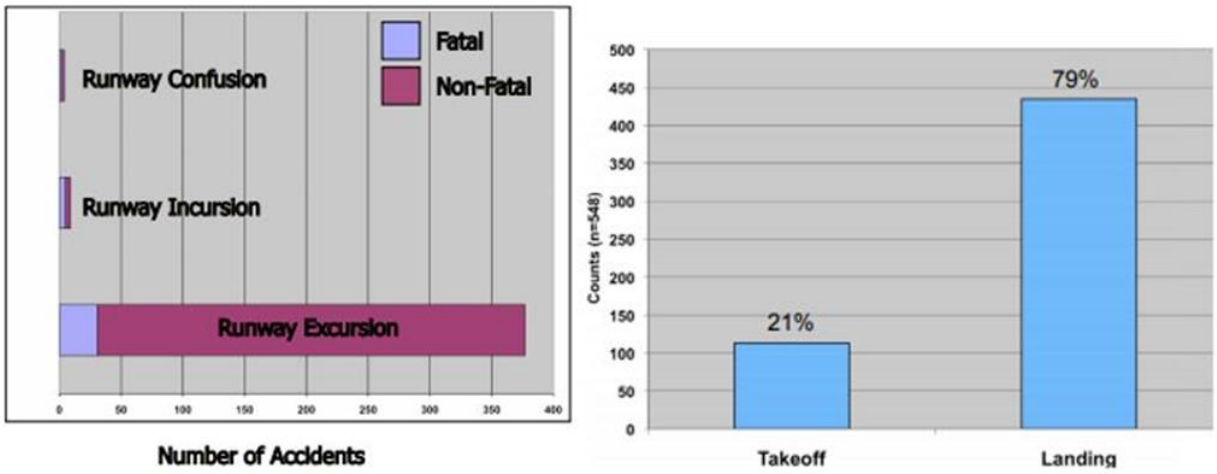
Nanduri (2016) used recurrent neural networks (RNN) with long short-term memory (LSTM) and gated recurrent unit (GRU) architectures on FOQA-like datasets to detect 11 canonical anomalies. RNN with LSTM and with GRU were used and Multiple Kernel Based Anomaly Detection (MKAD) was chosen as the anomaly detection method. It was concluded that all RNN

models are able to detect 9 out of 11 anomalous cases with no false positive. Janakiraman (2017) used a deep temporal multiple-instance learning (DT-MIL) model (weakly-supervised learning) to perform precursor mining for approach flights. High-speed exceedance (HSE) during landing phase was considered as the safety incidence and variables that are highly correlated to speed such as airspeed and ground speed were ignored.

Klein-Paste (2012) collected data to aid in Norway's initiative for an alternative operational decision support system to assess runway surface, where large and unique datasets correlating aircraft braking performance of commercial aircraft during “normal” operation on winter contaminated runways with detailed runway information reports and meteorological data were explored. Klein-Paste (AlexKlein-Paste, 2015) proposed a decision support model to help runway inspectors assess aircraft braking performance on winter contaminated runways. He showed that the model performed better than the predictions made by the runway inspectors or the friction measurements. Campbell (Campbell, 2016) developed a model validated using Global 5000 flight data to evaluate runway excursion based on the accuracy of the braking force in terms of reverse thrust. The uncertainty in the thrust reverser calculation is evaluated considering sensor error, and Monte Carlo simulation was used to determine how uncertainties propagate to the reverse thrust calculation. The results showed that the thrust reverser performance is mostly influenced by ambient temperatures.

2.3 Runway safety characteristics

Flight Safety Foundation published a report with a list of recommendations to flight operators, airport operators, air traffic management, and regulators regarding runway excursion mitigation (FSF, 2009). The report characterized runway excursion using worldwide statistical data from 1995 to 2008 for turbojet and turboprop aircraft as shown in Figure 3 (FSF, 2009), and quantified risk factors associated with takeoff and landing excursions. The report showed that determining whether any pair of associated factors has a causal connection would require a deeper study and analysis.



a) Number of Accidents Summary b) Accidents during Takeoff vs. Landing

Figure 3. Runway excursion worldwide statistics from 1995 to 2008

The European Aviation Safety Agency (EASA) published a report summarizing a list of precursor factors for several types of runway excursion accidents, including runway overrun after a rejected take-off (RTO), runway overrun on takeoff without RTO, runway veer-off on takeoff without RTO, runway veer-off after RTO, runway overrun after landing, and runway veer-off after landing (EASA, 2017). Precursors to each scenario were ranked by increasing "proximity" to the airport and by contribution to safety events. The report concluded with a list of recommendations to mitigate the impact of each precursor.

Jenkins (Jenkins, 2012) described a runway overrun mitigation tool developed by Boeing, called the runway mitigation situation awareness tool (RSAT) which goal is to determine runway overrun characteristics and potential causes as shown in Figure 4 (Jenkins, 2012). The tool further provides recommendations in order to mitigate overrun such as calculating the required runway length, determining the go-around point, and adding the thrust reverser call-out.

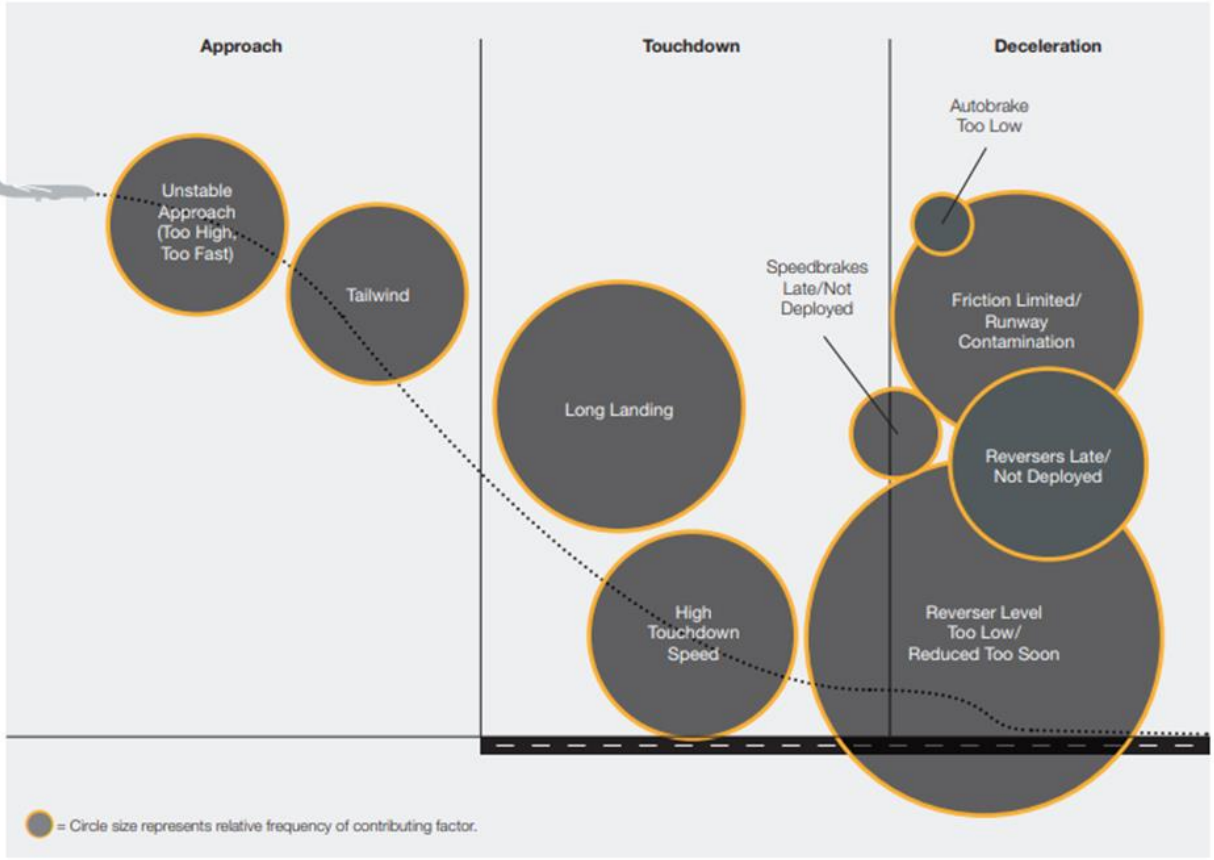


Figure 4. Causes of landing overrun excursions

2.4 Machine learning methods

Jasra (2018) conducted feature selection through unsupervised learning based on the single solution method, the multiple solution method, stochastic pruning, and node pruning. Then, anomaly detection was performed using unsupervised learning and distance-based, density-based, and probability-based approaches. It was shown that the density-based approach is more sensitive to detect flights with moderate to severe exceedance events compared to the probability-based method.

The results also showed that using a combination of both methods at a tight detection threshold is most likely to detect severe exceedance events compared to using either method alone. Memarzadeh (2021) developed a model based on convolutional variational auto encoders for expanding the labels of a sample data to the entire data population. The author performed a binary anomaly detection test (takeoff speed drop) and showed that, out of 16,000 samples,

100~1000 samples were labeled. He also performed a multi category classification test (normal, highspeed, path high, and flaps late classification for approach).

A comparison of the results based on precision and recall was conducted between the models built, namely Compact Clustering via Label Propagation (CCLP), two deep semi-supervised classification models (M1+M2 models), and Deep Temporal Multiple Instance Learning (DT-MIL). It was concluded that the CCLP and M1+M2 models performed similarly but better than DT-MIL, as shown in Figure 5 (Memarzadeh, 2021).

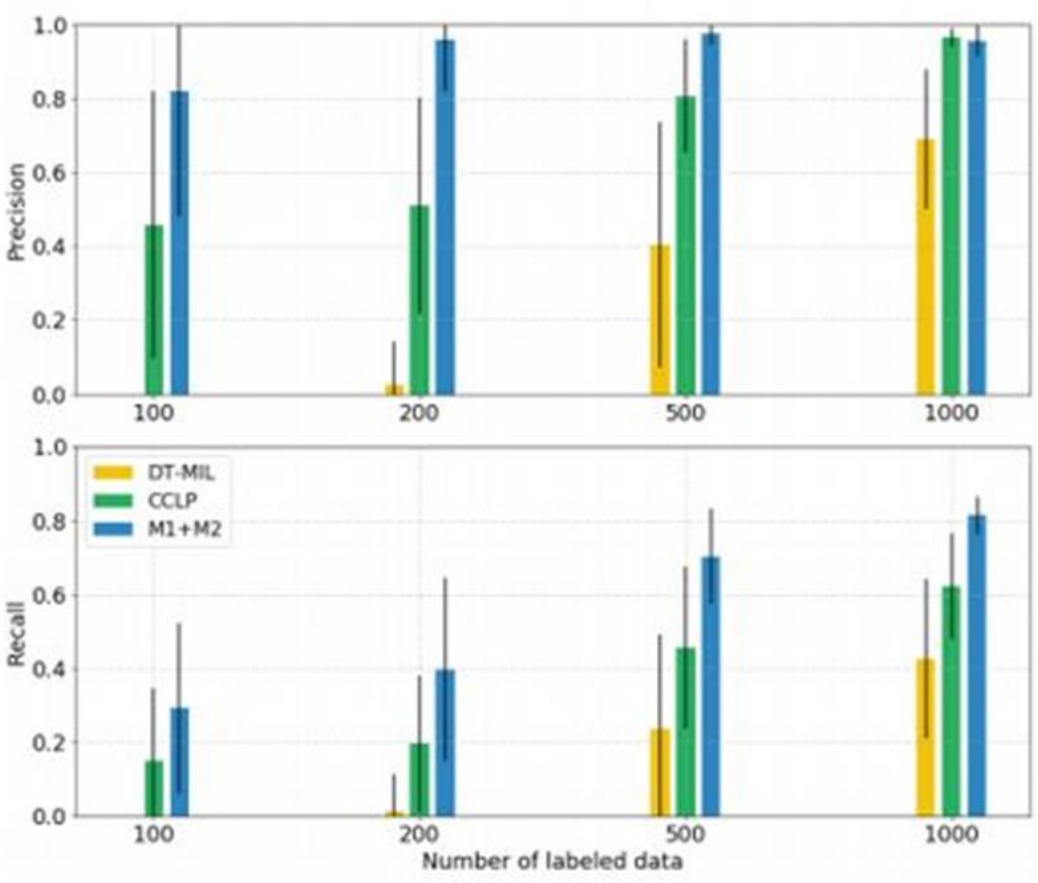


Figure 5. Performance of M1+M2 and CCLP vs. DT-MIL models on the binary anomaly detection task

Kirkland (2003) focused on properly modifying/normalizing the performance data for runway exceedance accidents. Data corrections are made based on the type of terrain, the aircraft performance, and the required distance to brake. The study derived wreckage location data relative to the end of the minimum required distance as well as the actual hard runway available

and allowed data for different types of aircraft to be consolidated into a dataset that has relevance for all aircraft types.

Jarry (2020) used a LSTM (long short-term memory) recurrent neural network model and simple dense models to estimate fuel flow rate, landing gear settings, and flaps configuration. The model predicted each metric for up to 70 nm with a normalized root mean square error (RMSE) below 10%. Li (2015) used clustering-based detection and multiple kernel anomaly detection methods to identify significant anomalies and concluded that the cluster-based model (ClusterAD-Flight) is able to categorize large numbers of flights quickly and performs better with continuous parameters. Zhang (2018) compared the performance of LSTM, SVM (support vector machine), BP (back propagating neural network), LoR (logistic regressions) to classify the prediction of a hard landing. He concluded that the LSTM model generated a prediction model with the highest F1 accuracy. Odisho (Odisho, 2020) used predictive models to determine pilot misperceptions of runway excursion risk associated with unstable approaches, and suggested that the ability to predict the probability of pilot misperception of runway excursion risk could influence the development of new pilot simulator training scenarios and strategies. Kang (2020) compared LSTM with common machine learning models such as decision tree (DT), linear regression (LiR), gradient boosting decision tree (GBDT), random forest (RF), neural network (NN), and SVM to generate a sequential ground speed prediction, both online and offline. The results are shown in Figure 6 (Kang, 2020).

Tong (2018) also compared LSTM, SVM, and Neural Network machine learning models to generate a sequential to point prediction of the ground speed, and performed a sensitivity study of the LSTM model based on the number of hidden layers. The author showed that the LSTM model allows an extremely accurate prediction of the metric of interest, but has a very short prediction interval that makes it hard to use for online applications.

Model	RMSE	MAE	MAPE (%)
GBDT LSTM Encoder-Decoder	**1.71**	**1.07**	**0.83**
Experience LSTM Encoder-Decoder	2.20	1.32	1.03
LSTM Tong et al.	1.96	1.50	1.13
Linear Regression	12.57	11.01	8.43
Decision Tree	5.70	4.66	3.51
GBDT	5.79	5.22	3.95
Random Forest	9.75	8.48	6.55
Neural Networks	21.86	21.17	16.11
SVM	7.83	6.55	5.02

Figure 6. Performance comparison between ML models to predict ground speed

2.5 Related works at the Aerospace Systems Design Laboratory

Puranik et al. (2020) evaluated the performance of the approach and landing phases of flight by choosing airspeed and ground speed as the significant parameters. A random forest regression algorithm was used to predict the most significant parameters to predict performance during approach and landing with an accuracy of RMSE below 2.62 and 2.98 knots.

The study confirmed that a global prediction model can be used to predict most significant parameters even when the model is generated from data collected from multiple airframes. Lee et al. (2020) developed a framework based on random forest classification to identify the ranking of significant parameters for a given flight event, and Tableau was utilized to identify anomalous events (such as “tire speed exceedance”).

The authors concluded that speed, weight, thrust, and air density have the highest contribution to the “tire speed exceedance” event. Other events such as “high roll” and “landing distant” were also studied. Mangortey et al. (2020) developed a framework to select a reduced set of significant parameters for a given event. The methodology consists of several steps, including a correlation analysis step to reduce duplicate parameters, a pre-processing step to remove parameters based on metadata and subject matter experts’ input, a clustering step to group similar flights and identify abnormal operations, a retrospective analysis step performed to

identify characteristics by cluster, and an Analysis Of Variance (ANOVA) to identify significant parameters for each cluster.

Ackley et al. (2020) created a framework based on sequential backward selection using random forest to identify precursors to flight events with multiple contributing factors. High frequency events were then identified for precursor identification evaluation. The significance of features was then aggregated from each feature vector as shown in Figure 7 (Ackley, 2020). The study concluded that energy management is a precursor for unstable approach events.

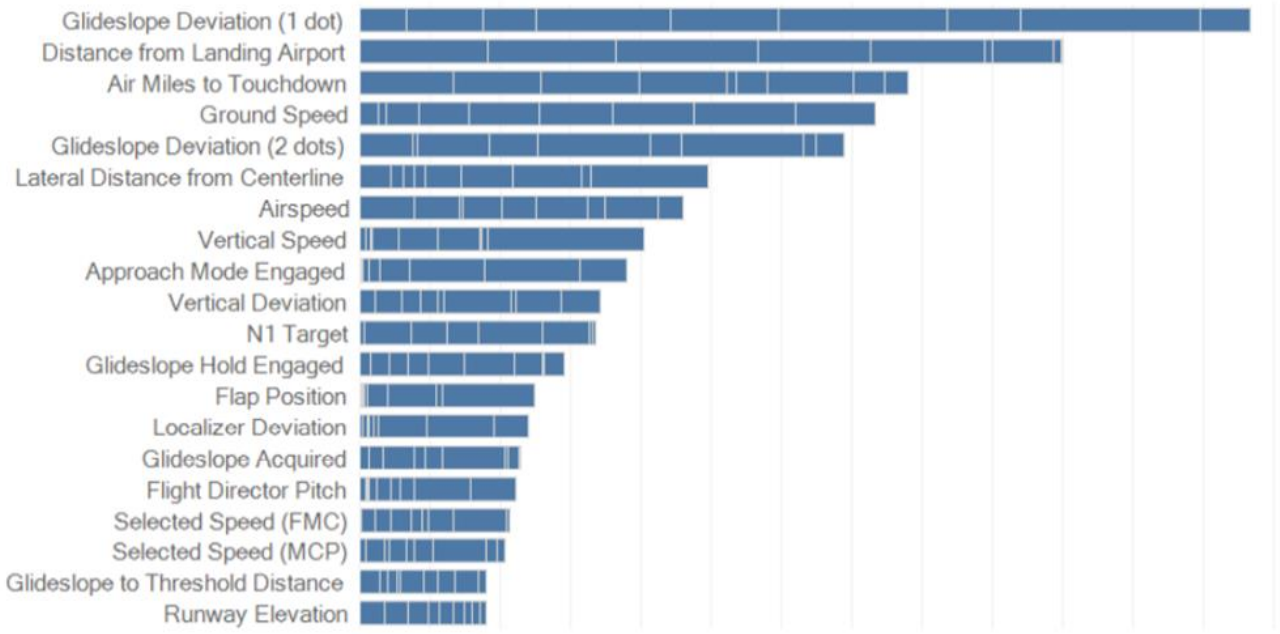


Figure 7. Aggregate cumulative feature importance

3 Data sources and fusion

To facilitate the data-driven analysis of degraded aircraft braking on contaminated runways, various sources of data are collected and processed.

3.1 Description and processing of data sources

This section describes in detail each data source utilized in the current research and the associated processing or cleaning.

3.1.1 Flight Operations Quality Assurance (FOQA) data

The first set of data of interest includes the metadata associated with flights. The metadata is unique to each flight and contains information such as aircraft type, date of operation, take-off and landing airport, airport elevations, runway length and width, runway slope, etc. The metadata can add context to the quantitative metrics of interest and also enables fusing or merging the flight data with other sources of data such as runway conditions, prevailing weather, etc.

The other set of data of interest consists of time series measurements obtained from on-board sensors and recorders. This data is recorded at frequencies of up to 16 Hz and may contain a large number of parameters. The various flight parameters may be organized in a hierarchical manner, going from groups of parameters to each individual parameter recorded in various units when possible. A notional illustration of a part of this hierarchical parameter tree is provided in Figure 8.

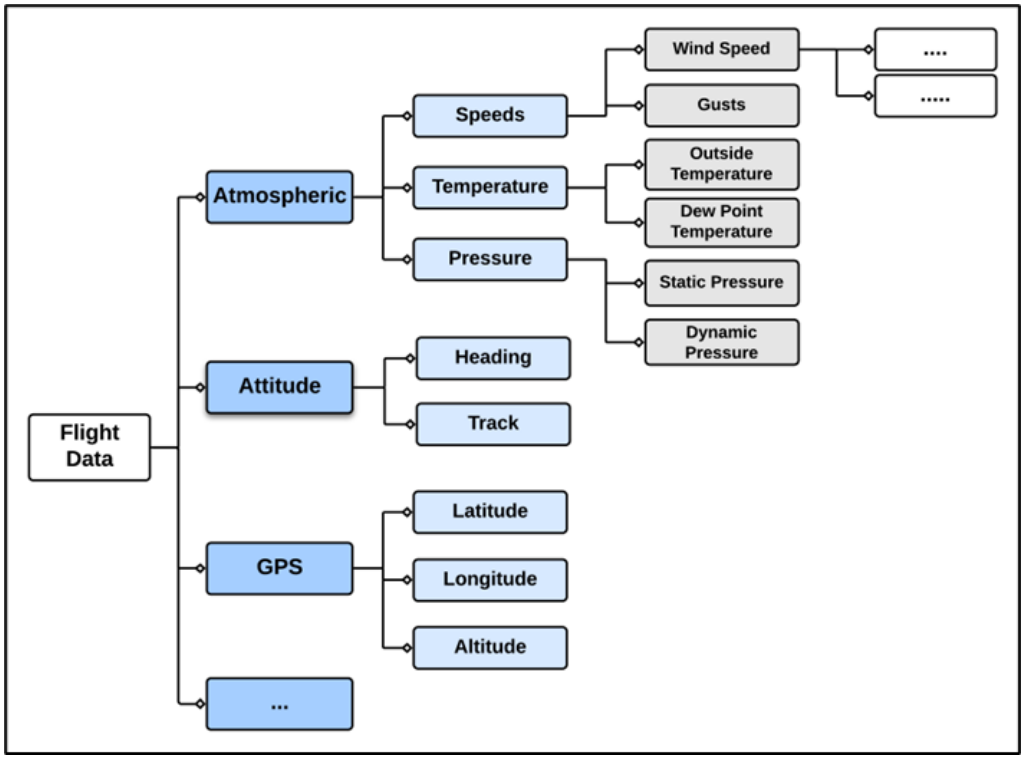


Figure 8. Notional Hierarchical breakdown of recorded flight data parameters

Only unique flight parameters (independent of their units) associated with the approach, landing, rollout, and taxi phases are retained for analysis in this study. Extracted flight parameters may be classified into the following six broad categories:

1. Structure Parameters
2. Aerodynamics Parameters
3. Operational Parameters
4. Ambient Conditions
5. Engine Parameters
6. Controls Parameters

There are other parameters that are also recorded or available that have not been used in this research and are therefore not discussed in this report.

Under the *Structure Parameters* category are brake pressure, normal force through gear, weight, wheel speed, and load factor. These parameters represent the stress experienced by the aircraft, wheels, and the landing gear system. For the landing and takeoff performance analysis, the normal force through the gear defines the type of dynamics experienced by the aircraft (air versus ground). It can also be used to identify when the aircraft has touched down on the ground along with parameters such as weight-on-wheels (WOW).

Under the *Aerodynamics Parameters* category are parameters such as drag, lift, angle of attack, pitch angle, etc. These parameters are representative of the aerodynamics performance of the aircraft and may have significant implications in the braking performance of the aircraft.

The *Operational Parameters* category contains parameters such as accelerations (normal, lateral, longitudinal), kinetic energy, potential energy, rate of descent, previous waypoint, altitude limits, etc. Some of these parameters represent the response of the aircraft to various actions by the cockpit crew while others represent the target state desired based on the aircraft configuration. As such, they are also used as threshold bounds that should not be exceeded during the operation of the aircraft.

Under the *Ambient Conditions* category are density, pressure, wind speed, and wind direction. These atmospheric parameters are external factors that cannot be controlled by the pilot but that have an impact on the aircraft performance. For example, wind speed and direction may define the aircraft heading during touchdown.

The *Engine Parameters* category is composed of parameters such as thrust, N1%, fuel flow rates, etc. These parameters are the direct result of the pilot's inputs, and they exist for each engine. Depending on the situation, aggregate or differential engine measurements may be used.

Under the *Controls Parameters* category are thrust reverser position, speed brake position, and various control surface positions (flaps, slats, etc.). These parameters record the pilot's inputs and may be further classified into *braking* devices and *control* devices. The parameters mentioned in this subsection do not represent the full list of parameters that are available, but the ones that are of importance for this research.

3.1.2 Weather data from ASOS/NOAA

Automated Surface Observing System (ASOS) units are automated sensor suites that are designed to serve meteorological and aviation observing needs. There are currently more than 900 ASOS stations in the United States with most of them located at or near airports. The system sensor data is publicly accessible through the official website of the National Oceanic and Atmosphere Administration (NOAA) National Centers for Environmental Information (NCEI)[2], which is “responsible for preserving, monitoring, assessing, and providing public access to the Nation’s treasure” of climate and historical weather data and information. The NOAA data repository provides ASOS weather data with both one-minute and five-minute intervals. In this research, the one-minute ASOS data has been chosen due to the rapid pace at which runway conditions and aircraft operations might deteriorate with adverse weather.

The NOAA weather data repository covers the time range from January 2000 to the most recent month and is split into two parts. The first part contains station ID, year, month, day, hour, minute (both local and Universal Time Coordinated (UTC)), visibility, extinction coefficient, speed of two-minute average wind (knots), direction of two-minute average wind (knots), speed of five-second average wind (knots), direction of five-second average wind (degrees), and runway visual range (hundreds of feet). The second part contains station ID, year, month, day, hour, minute (both local and UTC), precipitation amount (hundredths of inches), precipitation type, station pressure from three sensors (inches of Mercury), average one-minute dry bulb temperature (degrees Fahrenheit), and average one-minute dew point temperature (degrees Fahrenheit). Both parts of the information are stored as data files in a monthly .dat format, while the first part follows the name convention: "64050XXXXYYYYZZ.dat", and the second part follows the name convention: "64060XXXXYYYYZZ.dat", where XXXX is the four-digit International Civil Aviation Organization (ICAO) identifier for the ASOS station, YYYY is the

[2] National Centers for Environmental Information (NCEI) (noaa.gov)

year in two-digit format, ZZ is the month in two-digit format, and the leading four digits distinguish the file from containing the first part of the weather information or the second (6405 for the first and 6406 for the second).

3.1.3 Field Condition Report (FICON) data

In late 2016, the FAA alongside the TALPA ARC produced a new set of recommendations guiding aircraft performance and surface condition assessment and reporting. One of the most significant of these recommendations was the introduction of a consistent method for assessing runway conditions, known as the Runway Condition Assessment Matrix (RCAM). The proper application of the guidance in the RCAM is predicated on adhering to the limitation and assumption notes associated with the RCAM as found in AC 150-5200-30D (FAA, 2020). Figure 9 (FAA, 2020) displays the RCAM.

This matrix is visually divided into two sections, *Runway Assessment Criteria* and *Downgrade Assessment Criteria*. The Runway Assessment Criteria, applicable to paved runways (no turf, dirt, gravel, or waterways), provides airport operators the ability to connect runway contaminant types and depths to a Runway Condition Code (RwyCC). Airport operators may use the Downgrade Assessment Criteria, involving friction coefficient measurements, Pilot Reports (PIREPs), and their best judgement and experience to downgrade RwyCCs to a more conservative report.

The RwyCCs, along with runway specific information, are reported and distributed in Field Condition (FICON) Notices To Airmen (NOTAM). In the FICONs, a RwyCC value is reported for each third of the runway. Time is reported in the format: year, month, day, hour, minute. The general format for FICONs with italicized variables is as follows:

> !*Airport NOTAM_Number Airport Location Identifier* FICON *RwyCCs Contaminant_Type* OBSERVED AT *Observed_Time*. *Start_Time-Expiration_Time*

An example FICON follows:

> !*ADQ 01/492 ADQ RWY 01* FICON *5/5/5 100 PRCT WET* OBSERVED AT *1801312351*. *1801312351-1802012351*

From each NOTAM and metadata from the NOTAM Manager, the following metrics have been extracted: Airport, NOTAM Number, Runway, RwyCCs, Contaminant Description, Start Time, End Time, and Cancel Date/Time.

Assessment Criteria		Downgrade Assessment Criteria		
Runway Condition Description	**Code**	**Mu (μ) [1]**	**Vehicle Deceleration or Directional Control Observation**	**Pilot Reported Braking Action**
• Dry	6	40 or Higher	---	---
• Frost • Wet (Includes Damp and 1/8 inch depth or less of water) ***1/8 inch (3mm) depth or less of:*** • Slush • Dry Snow • Wet Snow	5		Braking deceleration is normal for the wheel braking effort applied AND directional control is normal.	Good
5° F (-15°C) and Colder outside air temperature: • Compacted Snow	4	39 to 30	Braking deceleration OR directional control is between Good and Medium.	Good to Medium
• Slippery When Wet (wet runway) • Dry Snow or Wet Snow (Any depth) over Compacted Snow ***Greater than 1/8 inch (3mm) depth of:*** • Dry Snow • Wet Snow ***Warmer than 5° F (-15°C) outside air temperature:*** • Compacted Snow	3		Braking deceleration is noticeably reduced for the wheel braking effort applied OR directional control is noticeably reduced.	Medium
Greater than 1/8 (3mm) inch depth of: • Water • Slush	2	29 to 21	Braking deceleration OR directional control is between Medium and Poor.	Medium to Poor
• Ice [2]	1		Braking deceleration is significantly reduced for the wheel braking effort applied OR directional control is significantly reduced.	Poor
• Wet Ice [2] • Slush over Ice [2] • Water over Compacted Snow [2] • Dry Snow or Wet Snow over Ice [2]	0	20 or Lower	Braking deceleration is minimal to non-existent for the wheel braking effort applied OR directional control is uncertain.	Nil

Figure 9. Runway condition assessment matrix

3.1.4 Runway and airport data

The FAA provides public access to a repository with airport and runway data that can be used for this research[3]. The data repository covers all Federal Aviation Regulation (FAR) 139 certified airports in the United States. This source is selected due to its high reliability and expansive coverage. Additionally, the repository is maintained by the FAA, so the data is up-to-date and includes a description of when it was last updated.

Five Microsoft Excel files are accessible in this database, four of them being data files and the other a description file. The four data files are airport facilities data, airport runways data, airport

[3] https://adip.faa.gov/agis/public/#/airportSearch/advanced

remarks data, and airport schedules data. The airport facilities data file contains basic information such as location, status, repair service availability, etc. The airport runway data file contains information such as runway ID, surface type and condition, runway treatment, runway end elevation, runway length and width, runway crossing height, etc. The airport remarks data file contains text data with other information about airports. The airport schedules data file contains the availability information about airports. Lastly, the description file, also known as the airport dictionary file, contains a detailed explanation regarding the four data files. Within the scope of this research, the airport facilities and airport runways files are of primary interest.

3.2 Data fusion pipeline

A data fusion pipeline has been built to integrate the heterogeneous data described in Section 3.1, namely FOQA flight data, ASOS/NOAA weather data, FICON data, and FAA runway and airport data. A consolidated database that covers all aspects of landing flights is produced by the pipeline and is used for building machine learning models. The data fusion pipeline is implemented using the Python programming language and its Pandas library.

Figure 10 depicts the overall process of the data fusion pipeline and specifies the content and the number of flights within each output file.

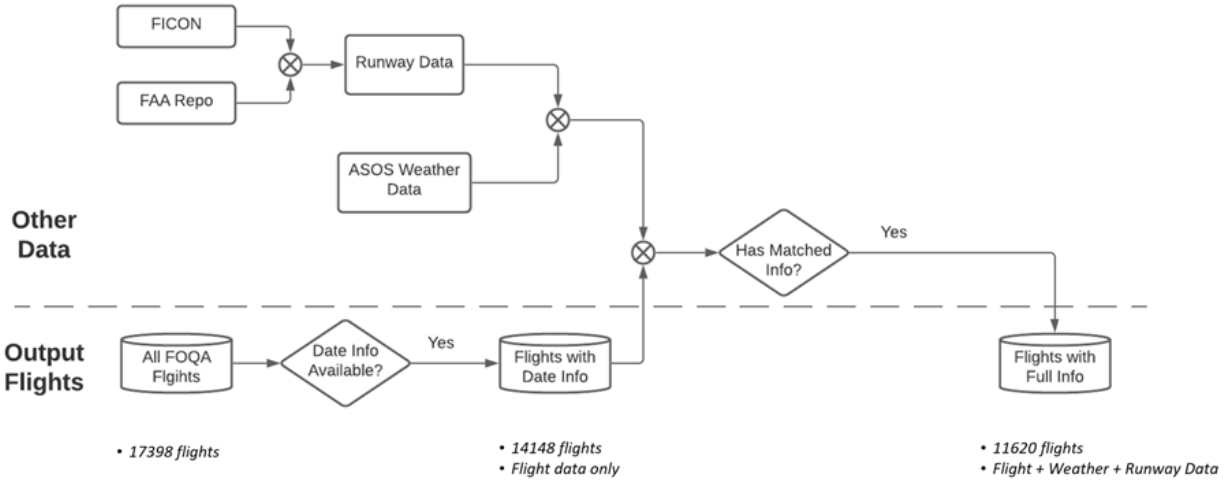


Figure 10. High-level data fusion pipeline

Figure 11 provides the details of the data fusion pipeline, such as the keys used for merging different data sources and expansion of time intervals.

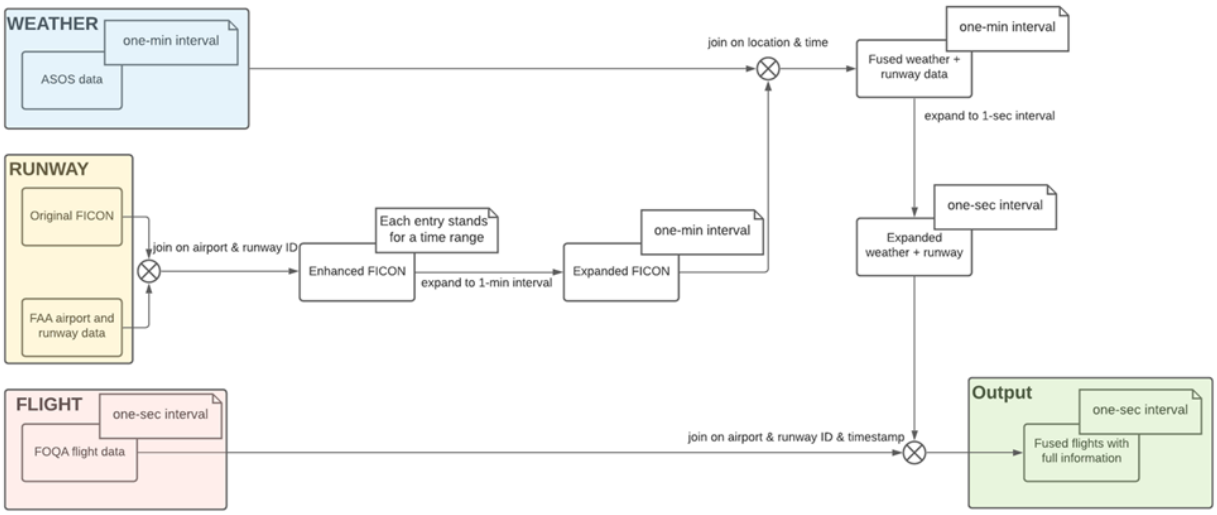


Figure 11. Details of data fusion pipeline

The final output of the data fusion pipeline is shown in Figure 12. The output file contains time series data with an interval of one second, wherein each row in the output represents the state of a particular flight (specified by the “Flight ID” column) at a particular second (specified by the “Time” column). Therefore, the entire landing procedure of a flight extends across multiple rows.

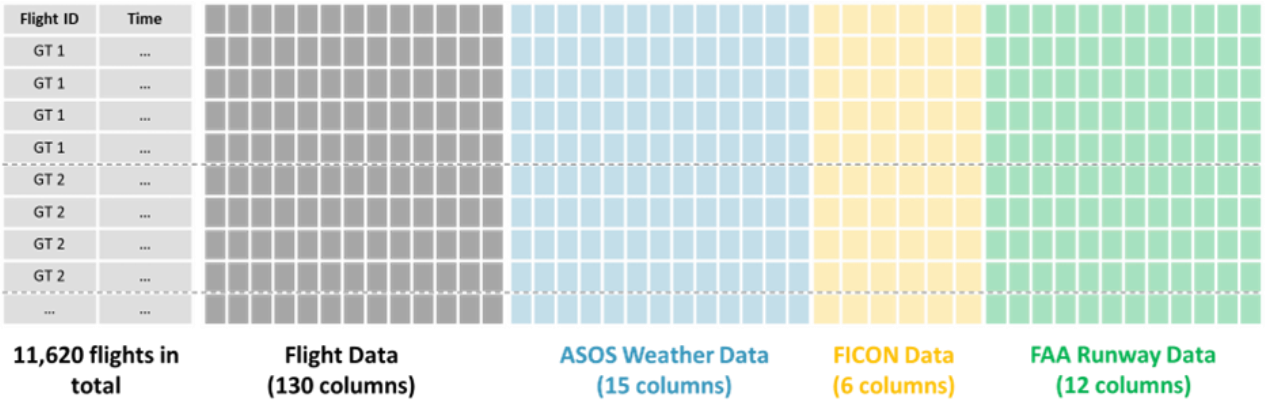


Figure 12. Data fusion final output

4 FICON data analysis

4.1 Dataset description

The data available for this study includes FICON reports from the winter months between the years 2016 and 2019. The weather conditions and airport and runway data for the corresponding time frames are obtained and fused. It is noted that during the time frames that the FICON reports are collected, a small proportion of the reports also contain the pilot reported braking action.

By default, in the dataset available, FICONs expire 24 hours after their effective time begins. However, designated observers can cancel or amend a FICON prior to the default expiration period. For the FICONs reported, 84.68% are cancelled and 15.32% expired.

4.1.1 Data distribution

In the fused data set, there are 683,145 rows and 25 columns of enhanced FICON data. Of this enhanced data, 568,791 rows (83.26%) contain RwyCCs and only 11,899 rows (1.75%) contain pilot reported braking actions. The intersection of these is 9,906 rows (1.45%) with RwyCCs and pilot reported braking actions. The data is largely distributed across the winter months, as seen in Figure 13(a). This is consistent with the expectation that the winter months are when degraded braking operations might be expected to occur.

4.1.2 Runway condition types

Runways with a consistent distribution of contaminants down the length of the runway being evaluated will typically have the same RwyCC for all three thirds (each RwyCC represents the

condition on one third of the runway). In this report, these will be referred to as Uniform FICONs. In the current dataset, as seen in Figure 13(b), nearly all (98.08%) FICONs are uniform. The remaining non-uniform FICONs have a dispersed distribution among the various non-uniform FICON combinations possible.

As observed from Figure 13, among the uniform FICONs, the majority contain the RwyCC 5/5/5 followed by 3/3/3 and 1/1/1. 4/4/4 and 2/2/2 make up the remaining small proportion of uniform FICONs. Any runway reported as "NIL" (meaning there is an empty data field) is closed for operations until the weather improves or contaminant removal is completed.

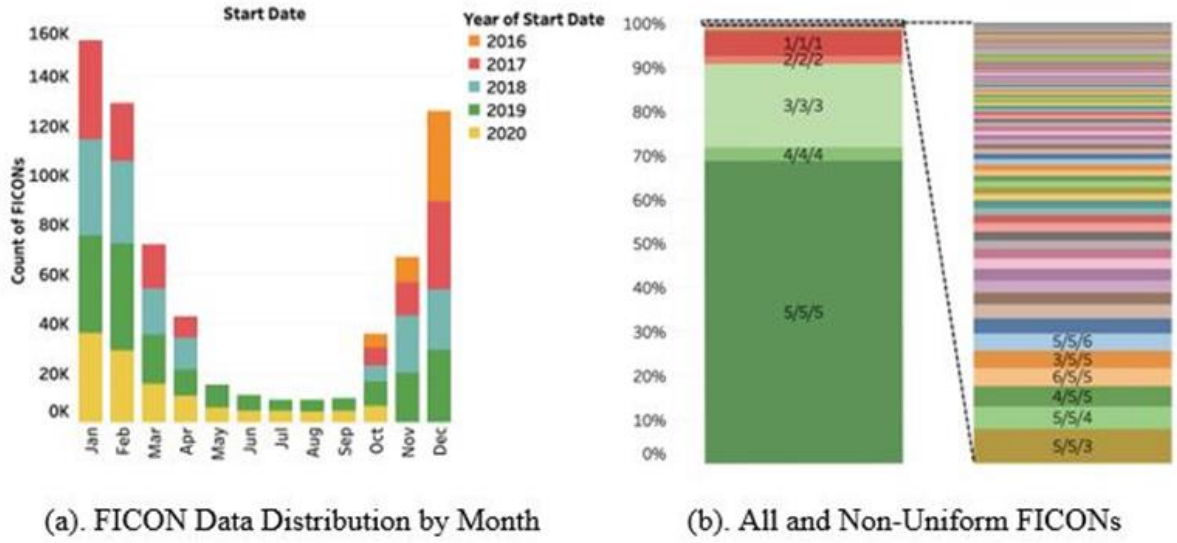


Figure 13. FICON data inspection

4.2 Univariate analyses with FICON data

In the enhanced FICON dataset, there are 1,951 distinct airports. Of these airports, 976 (50.1%) are located within the United States. The airports located domestically account for 613,448 of the 682,757 FICON reports with locations (89.85%). The remainder of this subsection contains the results of the analysis of the enhanced fused datasets and their implications on runway safety. It is divided into two parts, one subset which contains RwyCCs amounting to 568,791 samples, and the other subset which contains the pilot reported braking actions (PIREP BA) amounting to 9,906 samples.

4.2.1 Subset 1

This subset provides a basis for making a correlation between the physical construction of a runway and its associated surface characteristics. To describe the construction characteristics of a runway, three descriptor categories are commonly used in the regulatory and research literature:

1. *Runway treatment*: refers to modifications to the surface to reduce standing water and hydroplaning potential (Es G. V., 2001). Modifications may be: Grooved (GRVD), no treatment (NON), and porous friction course (PFC).
2. *Runway Surface Type*: describes the material used in the runway construction, which may be: Asphalt, Asphalt/Concrete, and Concrete.
3. *Runway Condition*: refers to the quality of the runway surface and is an indication of proper maintenance by the airport operator. Condition may be: Excellent, Good, Fair, Poor.

The detailed impact of *Runway Condition* on aircraft braking performance is deferred to future investigations. *Runway Surface Type* is quantified for reference, but our main analysis focuses on the influence of *Runway Treatment* on aircraft braking performance. These two attributes are visualized in Figure 14(a).

From the larger dataset in Figure 14(a), a down-selection is conducted to evaluate only events that contain both a *Runway Treatment* value and a reported RwyCC. The results are shown in Figure 14(b). This figure illustrates that there is a significant trend correlation between the type of runway treatment and the history of RwyCCs reported. It can be observed that larger values of the RwyCC (RwyCC >= 3) are more commonly reported for *treated* runway surfaces (PFC and GRVD) compared to *no treatment* runway surfaces (NON). Since larger RwyCCs correspond to better runway conditions, this indicates that having runway treatment would improve runway conditions.

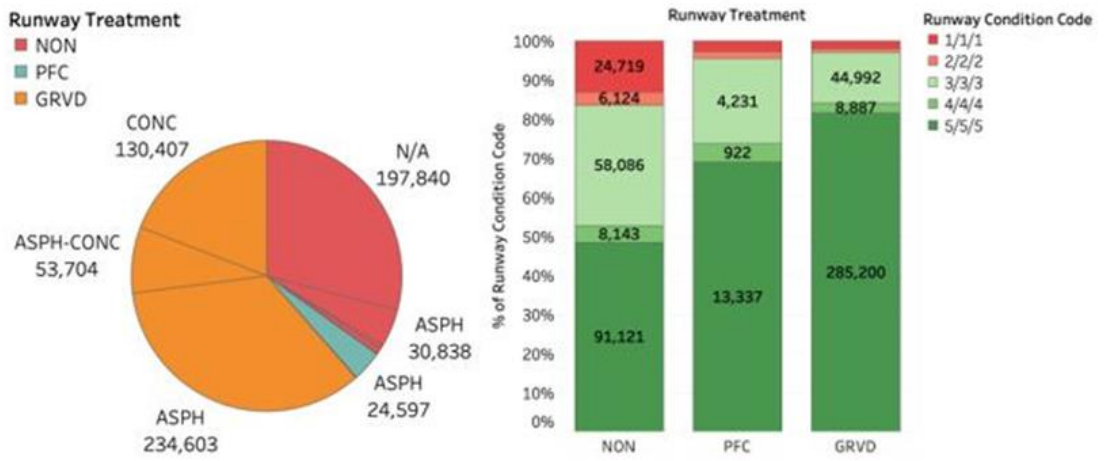


Figure 14. FICON subset 1 visualization

4.2.2 Subset 2

Using current "time of arrival" best practices guidance, one of the most important sources of braking action information for cockpit crews is from the reported RwyCCs. The RwyCCs are derived from the RCAM guidance using the runway contaminant type and depth observed at a limited number of sampling points on the airport property. While contaminant descriptors may be the primary determinant in the braking action reported, other static variables such as runway longitudinal slope, polished/rutted wheel tracks, or runway lateral slope (crown) may impact the braking action achieved. The sum of the RwyCC braking action plus runway variances should be reasonably reflected in the PIREP BA reports. This subsection presents the comparison of the reported RwyCC with the reported PIREP BA for the purpose of validating expected versus actual braking action.

- Reported braking action for contaminants and FICONs

  FICONs can contain both a RwyCC and additional descriptive text of a contaminate or level of coverage. To validate the consistency of this reporting, several visualizations are created. As seen in Figure 15(a), the expected trend of fewer GOOD / GOOD-MEDIUM / MEDIUM PIREP reports for the lower RwyCCs is confirmed. However, this does raise the question of why there are so many reports of GOOD even when the FICON reports 1/1/1 (POOR). It could perhaps indicate that the RwyCCs might be overly conservative or the PIREPs are overly optimistic. The data provides no clear basis for this apparent

bias. This points towards the need for potential additional criteria to be developed to understand and isolate these variations.

Figure 15(b) provides the closest correlation between contaminate descriptor and PIREP BA report, but with approximately 15 % of landings on ICE reported as MEDIUM or better, some questions remain as to the source(s) of the variance.

- Pilot braking action and non-contaminant variables

  This subsection analyzes the way the pilot reported braking action interacts with different variables for different runways. In the enhanced FICONs dataset, information regarding runway length, start elevation, and end elevation is available. This data and the following formula are used to calculate the longitudinal slope of each runway in the dataset.

$$slope = \frac{elevation_{end} - elevation_{start}}{length} * 100 \quad (1)$$

In Figure 15(c), runway slopes are grouped into bins of 0.1 degrees and plotted against pilot reported braking action. Although the domain of the chart includes [-0.3, 0.3] degree slopes, the segment [-0.2, 0.2] degree slopes is where most data is available, as runway slopes tend to follow a binomial distribution around 0. More positive slopes (uphill) have a strong positive relationship with 'good' reported braking action. This is best explained by the fact that positive slopes result in a force of gravity against the direction of motion.

From Figure 15(c), runway slope bins of -0.2, 0.0, and 0.2 degree have "good" braking action reports of 49.9%, 57.1%, and 68.4% respectively. As such, we find that there is a respective -12.6% and 19.8% change of "good" braking action reports for slopes of -0.2 and 0.2 degree in comparison to level ground.

Lastly, as referenced in the discussion of Figure 15(c), Figure 14(b) also shows the correlation between *runway treatment* and *braking action* and provides a validation of the RwyCCs.

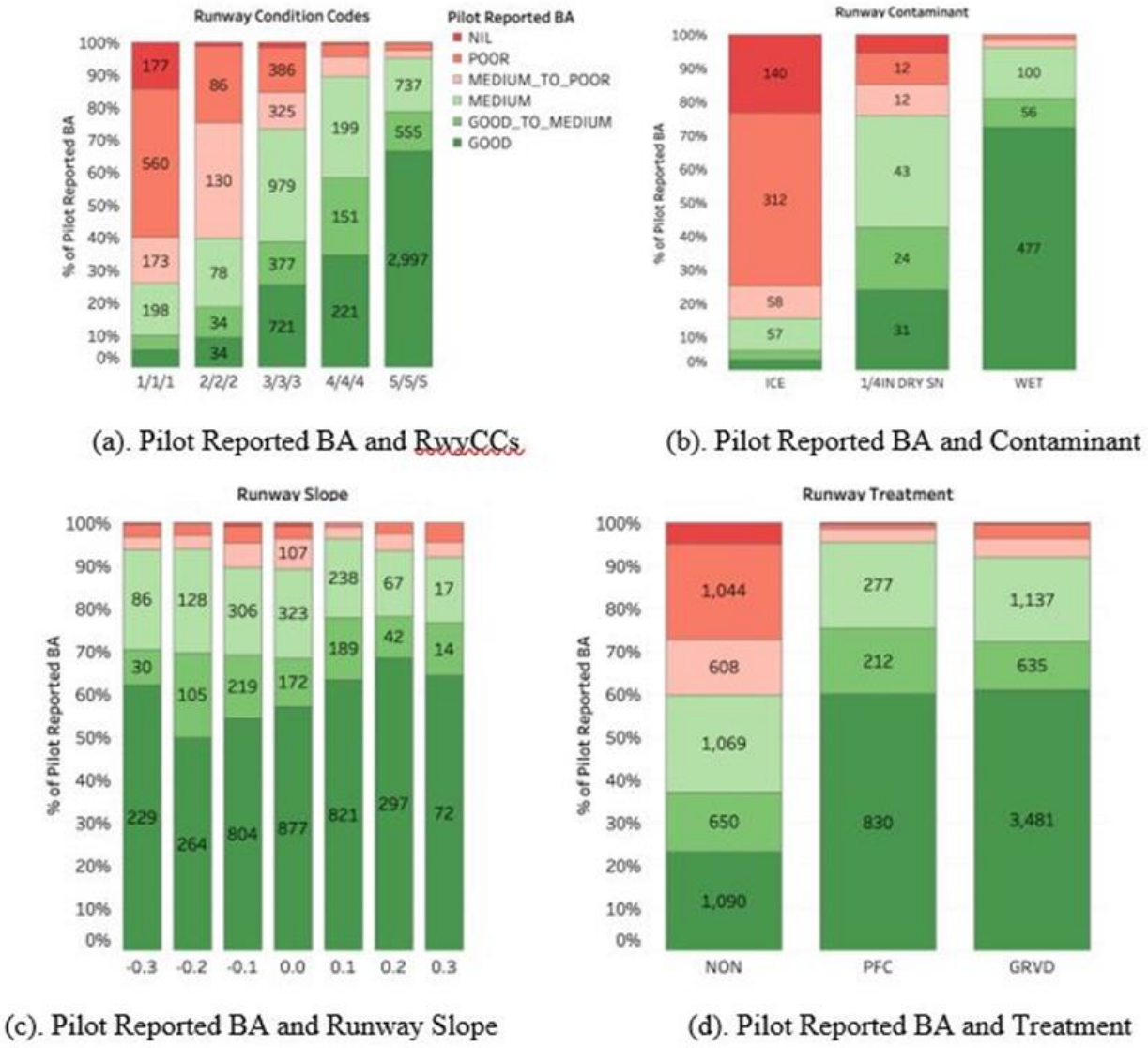


Figure 15. FICON Subset 2 visualization

4.3 Multivariate analyses with FICON data

To quantitatively explore the multivariable relationship between runway-related metrics and PIREPs, a series of analysis of variation (ANOVA) tests have been conducted. One-way ANOVA is typically used to investigate whether variations of a single factor have a measurable effect on a dependent variable[4]. Similarly, N-way ANOVA can be used to determine if there is

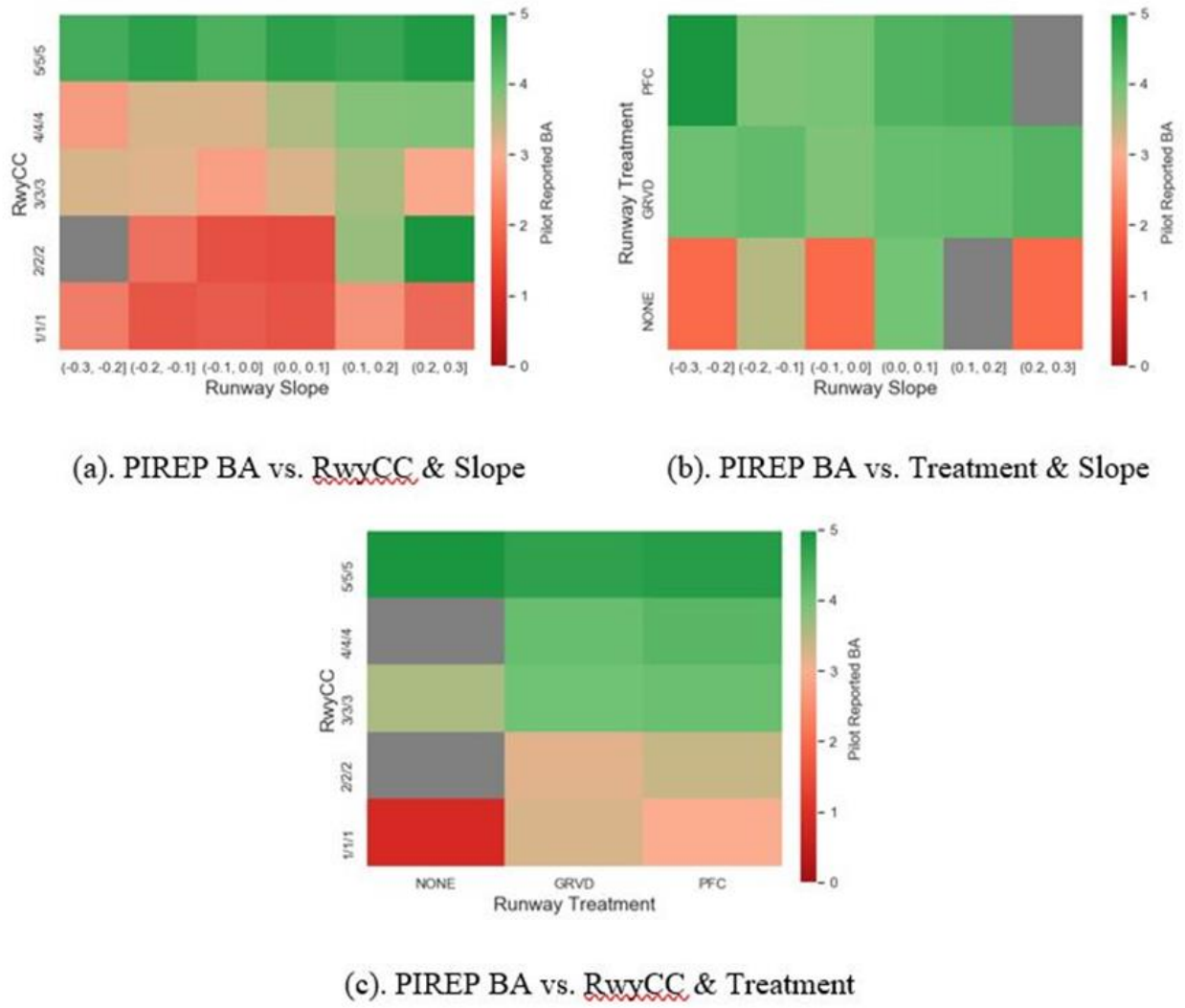
[4] One-Way ANOVA | Introduction to Statistics | JMP

an interaction effect between *n* independent variables on a continuous dependent variable (IGGA, 2009).

The runway-related variables are runway condition code, treatment, and slope. A subset of the dataset is selected for the analyses. It consists of 4,420 samples and contains full information about the three runway-related variables and pilot reported braking actions.

ANOVA requires the dependent variable to be numerical, so the *text* contents of PIREP BA are replaced by numbers from 0 to 5, where larger numbers correspond to more positive reports. For instance, “NIL”, which means no braking, is represented by 0, while “GOOD”, which means good braking performance, is represented by 5. Additionally, ANOVA requires the independent variables to be categorical, so runway slope has been converted to such a variable by grouping the runway slope values into bins of 0.1-degree width.

Figure 16(a) to Figure 16(c) depict the heatmaps of Enumerated Pilot Reported BA against different combinations of RwyCC, Runway Slope, and Runway Treatment. An individual cell in a heatmap represents the mean value of enumerated PIREP BA for a specific combination of runway variables. High PIREP BA values are represented by green while low PIREP BA values are represented by the color red. The grey cells in the heatmaps indicate there is no sample for the given combination.



(a). PIREP BA vs. RwyCC & Slope

(b). PIREP BA vs. Treatment & Slope

(c). PIREP BA vs. RwyCC & Treatment

Figure 16. Heatmaps for PIREP BA vs. runway variables (slope and treatment)

Three-way, two-way, one-way ANOVA tests are performed, and the results are summarized in Table 1. After looking at the reduced model which includes all possible 2-factor interactions, the only significant interaction found is between runway condition code and runway slope ($p <= 0.01$). One-way ANOVA tests indicate that runway condition code and slope have very significant effects on braking action ($p<= 0.01$), while runway treatment only has a significant effect on braking action ($p <= 0.05$).

Table 1. Results of ANOVA tests

	SUM_SQ	**DF**	**F**	**P**
RwyCC	632.95	4	153.45	< **0.01**
Treatment	10.06	2	4.88	< **0.05**
Slope	35.19	5	6.83	< **0.01**
RwyCC + Treatment	0.07	8	0.01	0.93
RwyCC + Slope	63.40	20	3.07	< **0.01**
Treatment + Slope	0.09	10	0.01	0.93
RwyCC + Treatment + Slope	54.27	40	1.34	0.26

Post-hoc tests (Tukey HSD[5]) are performed to further investigate the effects of the three runway variables and the interaction between RwyCC and Slope on braking action. These tests yield the following statistical results:

1. PIREP BA is significantly higher when RwyCC is higher.
2. PIREP BA is significantly higher when Treatment is PFC or GRVD compared to NONE.
3. PIREP BA is significantly higher when Slope has a larger positive value.

The tests also reveal the statistically significant interaction of RwyCC and Slope on PIREP BA. When Runway Slope is around zero (that is, -0.1 < Slope < 0.1), RwyCC has a significant effect on PIREP BA. However, when Runway Slope has a large absolute value, the correlation of RwyCC with PIREP BA becomes less robust. This observation can be explained by the fact that as the runway becomes steeper, the force of gravity starts to have more effect on the braking action, and the effect of RwyCC is diminished.

4.4 Correlation study with FICON and weather data

Runway braking conditions are affected by weather, specifically during rain or snow events. Quantifying the correlation between active weather events and reported braking action is the primary benefit of this investigation. It is acknowledged that reduced braking action reports unrelated to active weather are also possible, most notably for cold climate airports where compacted snow or ice may remain on a runway surface for weeks or months during the winter season (Klein-Paste, 2012).

[5] https://methods.sagepub.com/reference/encyc-of-research-design/n478.xml

The correlation between weather and RwyCCs is investigated and an example is presented below to give more details. The example covers all FICON records and weather data for Bangor International Airport in Maine (ICAO: KBGR) during the time range from 1/1/2018 to 12/31/2019.

Although a FICON record consists of 3 RwyCCs for the first, middle, and last thirds of the runway, most of the FICON records have uniform codes. As a result, the RwyCC representing the minimum of the ones having non-uniform codes will be used as a representative for a FICON record in our analysis. Also, while each FICON record covers a time span, only the start timestamp is used to represent the time of the record in the analysis. The correlation analysis of weather and FICONs addresses the relationship between temperature, precipitation type, and RwyCCs. In this report, all precipitation types that can lead to non-dry runway conditions (rain, snow, sleet) or dry runway conditions (no precipitation) are being considered.

Figure 17 displays the distribution of RwyCC with respect to temperature (air temperature and dew point temperature). In the ASOS system, 'R' stands for rain, 'S' stands for snow, and 'NP' stands for no precipitation, whereas '+' and '-' stand for heavy and light precipitation intensities. Each subplot in the relational plot shows the distribution of codes under one specific precipitation type. Within each subplot, the lower left part can be considered as the 'low temperature area' and the top right part can be considered as the 'high temperature area'. Each dot in the plot represents one FICON record, with its lowest RwyCC indicated by a color gradient: greenish colors represent better runway conditions, while reddish colors represent worse conditions.

Figure 17 shows that dots are spread over the 'low temperature area' under *snow precipitation type*, over the 'high temperature area' under *rain precipitation type*, and across the entire temperature area when there is *no precipitation*. This indicates that the temperature is usually low during snowy days and relatively higher during rainy days, and no precipitation can occur regardless of the temperature of the day. Moreover, several red dots are seen under snow condition, while green dots are the majority under *rain condition*, which indicates that snow can be more detrimental to runway condition compared to rain. Lastly, in the *no precipitation* subplot, it can be observed that most red dots appear in the 'low temperature area', which suggests that runway condition can be poor for temperatures near or below freezing even when there is no precipitation. Indeed, there are many other factors (besides contaminants) that can come into play for poor runway conditions (such as time of the day, dew conditions, tire 'hardness' vs. temperature, etc.). These observations, while straightforward, corroborate the

expected trends of correlations between adverse weather conditions and runway condition codes at a high level.

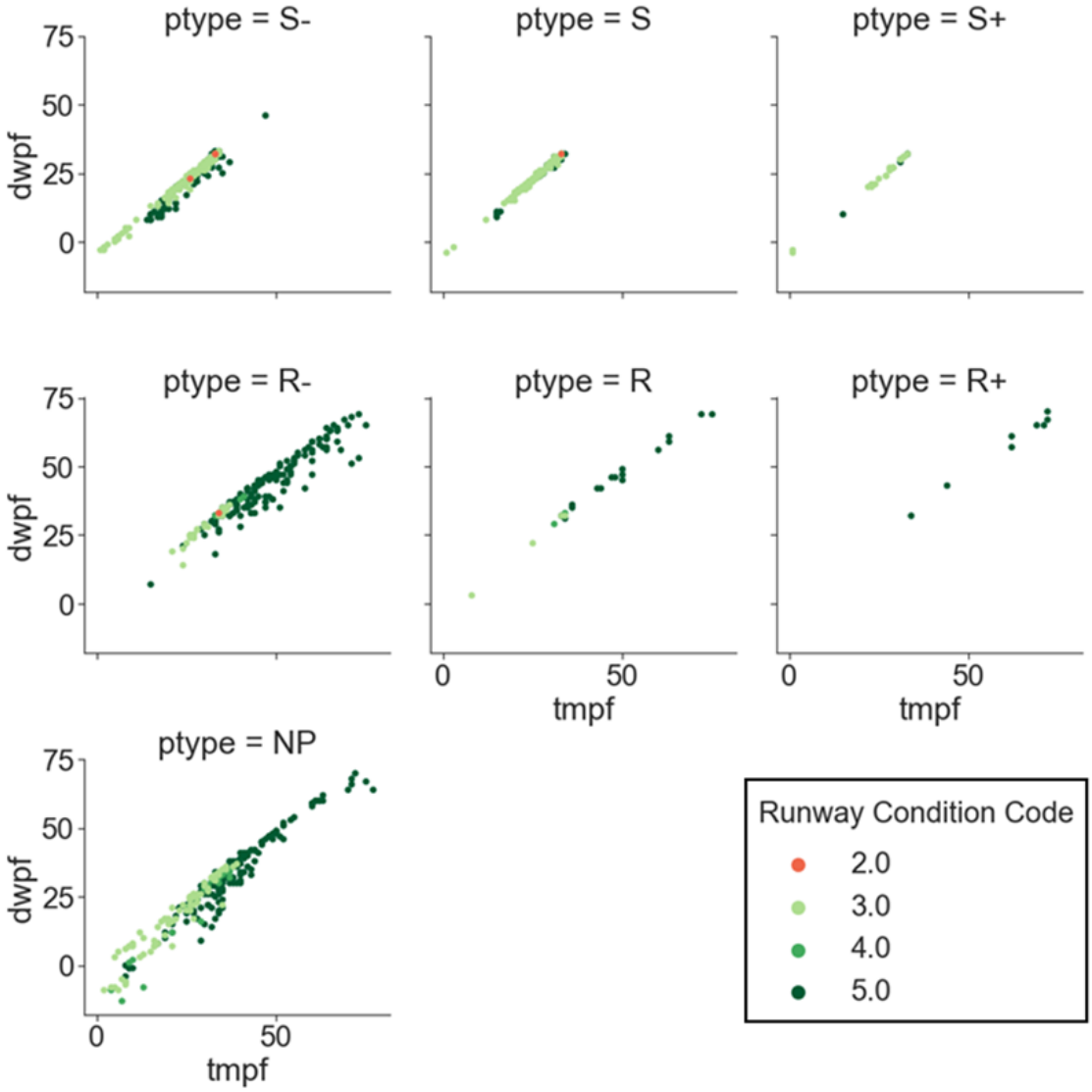


Figure 17. Relational plot of weather and runway condition codes at KBGR in 2018-2019

4.5 Conclusions

The results presented in the previous sections provide statistical support for the accuracy of existing best practices predictions for aircraft “time of arrival” braking effectiveness. Overall, the intention of this research is to present and collate available data in an understandable and functional manner to be useful for future studies.

The data also suggests that while the overall accuracy is reasonable, precision is more problematic. In aviation, conservative solutions are an essential part of safety. However, the RwyCC predictions appear to be significantly skewed towards a conservative performance level when compared to PIREPs during actual operations. The analysis of large datasets as performed in this research, may provide the only method to effectively reveal improved precision for performance predictions.

In this research, the relationship between runway surface conditions, airport and runway characteristics, prevailing weather conditions, and pilot reported braking action are studied over a large period of time using collected data. A robust and repeatable data fusion framework is developed to integrate data from various sources in order to analyze aircraft braking performance on contaminated runways. A statistical analysis is conducted to study the effect of prevailing weather conditions, runway treatment and slope, contaminant types, and other factors on the pilot reported braking actions and the runway condition codes.

The developed data fusion framework and FICONs are intended to be used in conjunction with real-world flight data. The eventual aim of the project is to be able to understand and infer runway conditions based on the collected and processed data using big data/machine learning techniques (Sheridan, et al., 2020; Mangortey, et al., 2020).

5 Reverse thrust analysis

5.1 Task and approach

Aircraft use multiple braking devices to maximize braking potential. These devices include thrust reversers, spoilers, wheel braking, and aerodynamic drag. Aerodynamic drag and spoilers contribute to aircraft deceleration by generating friction at the aircraft surface. Wheel brakes generate friction at the main gear to decelerate the aircraft. Thrust reversers generate thrust in the direction of motion to decelerate the aircraft. The way thrust reversers apply braking is different from other devices in that there is a delay between zero and full application due to the engine spool-down time. For this reason, the engine spool-down time for thrust reverser application needs to be modeled based on existing flight data to accurately model the real thrust reverser behavior in a simulation environment. The goal of this section is to model the spool-down time of thrust reversers until max braking application.

Flight Operations Quality Assurance (FOQA) data collected between the years 2013 and 2019 is used for this study. There is a total of 19,500 flight data categorized into 9 flight phases. For proprietary purposes, airframes are categorized as Narrow-Body (or single aisle) denoted NB,

and Wide-Body (or twin-aisle) denoted WB, and are numbered depending on their type and variant. For example, NB-B1/1 is narrow-body B1, variant 1. Data for the NB-B2/2 airframe with CFM International[6] engines is selected to narrow the data scope. This airframe and engine combination has the highest amount of data available for analysis. Also, this study focuses only on the approach and rollout phases of flight since these phases have a higher relevance to thrust reverser application. In terms of data parameters, the FOQA data contains over 2,000 parameters per flight. The relevant parameters are selected and listed in Table 2.

Table 2. FOQA parameter of interest for reverse thrust analysis

Thrust Lever Angle	N1 %	Reported Thrust
Thrust Reverser Position	Ground Speed	Height Above Touchdown
Ground Track Position	Latitude	Longitude
Corrected Gross Weight	Speed Brake Position	True Airspeed
Wind Speed	Wind Direction	Brake Pressure
Flight ID	City Pair	Acceleration
MSL Altitude	Lateral Distance from Runway Centerline	

These parameters are chosen based on their availability for the approach and rollout flight phases. They are time-variant quantitative data since most modeling algorithms require quantitative data for training. The data types include continuous, discrete, and binary types. The recording frequency of the FOQA data ranges between 0.25 to 16 Hz. For this study, the 1 Hz data frequency is used across all parameters since the minimum recording frequency of the selected parameters is 1 Hz. Note that flight ID and city pairs are metadata that are not time-variant, so they are repeated for the full data length for each flight. Some empty data points are observed due to errors in recording in real-time. These empty cells are kept empty. Some parameters such as *ground track position over full runway length* are augmented based on the existing data to better represent the braking progress. Braking control parameters, such as *% runway position at thrust reverser application*, are augmented so that zero represents extremely late use of thrust reverser or no use of thrust reverser.

[6] See: About CFM - CFM International Jet Engines CFM International (cfmaeroengines.com)

The data is compiled such that each column represents the recording of a specific parameter and each row represents a timestamp of any given flight. An illustration of this structure is shown in Table 3.

Table 3. Illustration of data structure for each flight thrust reverser analysis

Offset	Arpt Elevation (ft)	Arpt Latitude (degrees)	Arpt Longitude (degrees)	EGT left (deg C)	EGT right (deg C)	N1 left (%)	N1 right (%)	N2 left (%)
15571	1026.2	33.6367	-84.4279	476.5	469	47.71875	47.125	75.1875
15572	1026.2	33.6367	-84.4279	476.5	471.5	47.8125	46.96875	75.40625
15573	1026.2	33.6367	-84.4279	471	468.5	47.96875	47.34375	75.21875
15574	1026.2	33.6367	-84.4279	467.5	465.5	47.71875	46.9375	74.78125
15575	1026.2	33.6367	-84.4279	467.5	466.5	47.5625	47.125	74.875
15576	1026.2	33.6367	-84.4279	468	462.5	47.625	47.125	75.125

The data flow is illustrated in Figure 18. The process starts by combining Meta and FOQA data used for the thrust reverser modeling. This process includes augmenting the metadata to the length of each approach and rollout phase of flight, then concatenating the data horizontally. The combined data is then used to augment new parameters that represent the ground track position and braking progress. Augmented parameters include progress for full approach and rollout, progress from 50 ft altitude to end of the rollout (when the aircraft exits the runway), and temporal progress for full approach with timestamp (= 0 at touchdown).

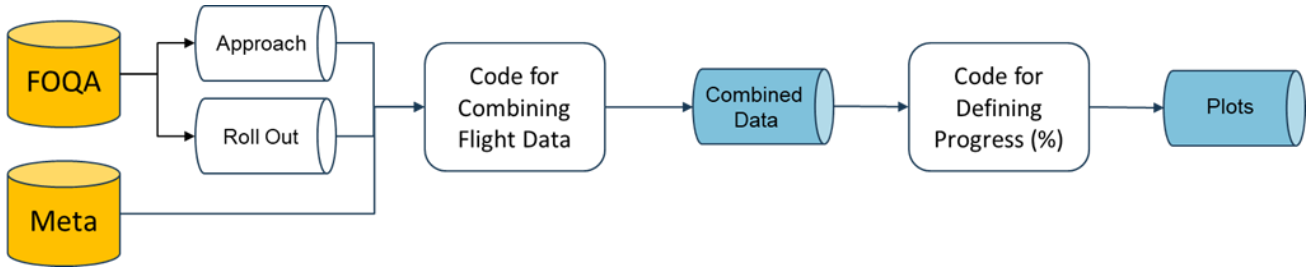


Figure 18. Data flow for thrust reverser spool-down time modeling

Samples of the modified data are depicted in Figure 19 to show the overall trend in the touchdown and braking behavior.

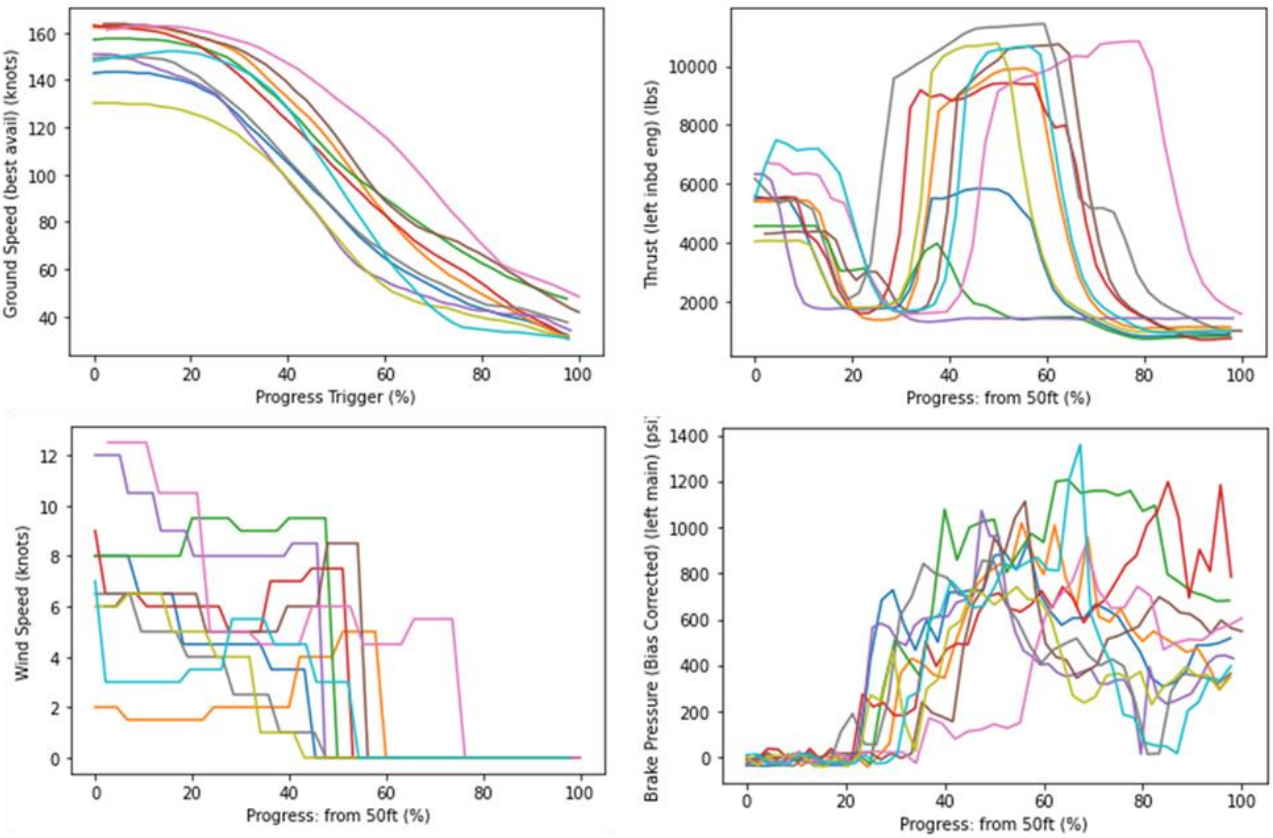


Figure 19. Observation of ground speed, wind speed, and various braking devices application

Figure 20 shows that, for a typical flight when the thrust reverser is used, the N1 % parameter goes up to 70%. Brake pressure data contains a lot of noise. The point of braking application may however be identified due to the fact that before brake application the recorded, pressure is close to zero with some noise. The braking devices are applied approximately at 20% from 50 ft height above touchdown to runway exit. Significant ground speed deceleration is observed during the ground roll where thrust reverser and wheel braking are applied, which is expected. Wind speed is an external factor that may influence the braking performance, but a simple observation of the wind speed and ground speed trajectory does not show a significant relationship between the two parameters.

To get a deeper understanding of the flight data, recordings of aircraft control parameters are observed and compared. An illustration of control parameters over ground roll progress are shown in Figure 20.

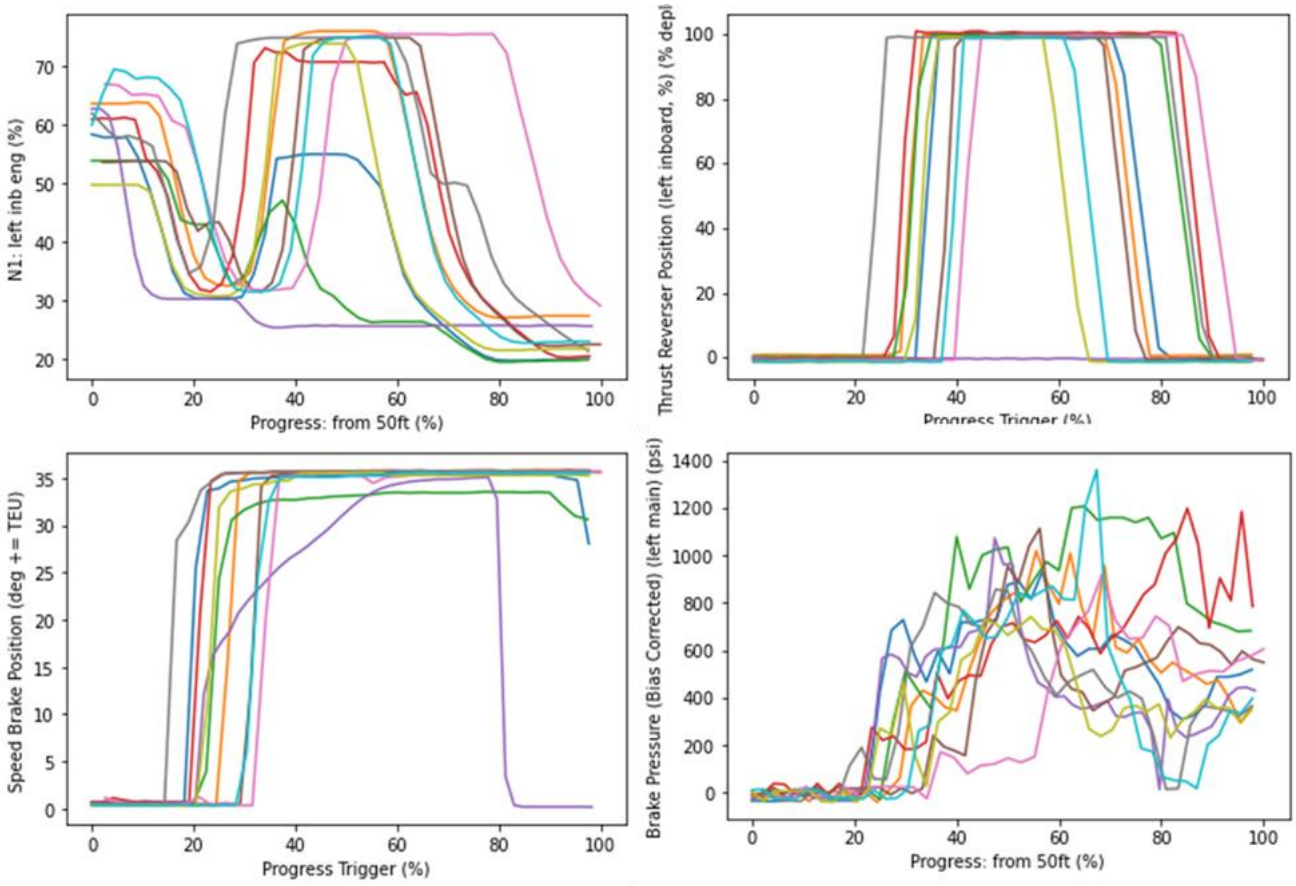


Figure 20. Observation of braking control parameters during braking rollout

Simple observation shows that the thrust reverser surface deflection and N1 % typically deploy simultaneously. However, this does not mean the two are turned off simultaneously. The same applies to the spoiler and the brake pressure turn off where these braking devices are applied at a similar time as thrust reversers. However, for a typical flight, the spoiler remains deflected and brake pressure remains applied past the end of the rollout.

The data is further analyzed to observe the thrust reverser spool-down behavior. The frequency of minimum N1 % and maximum N1 % from touchdown is observed to determine how the spool-down needs to be modeled. This data was collected from the final time index when 95% of the maximum N1% is reached (referred to as t1) to the first time when 105% of the minimum N1% is reached from t1, for each flight. This way the ranges of N1 % and time that need to be modeled for the spool-down behavior are determined. As shown in Figure 21, approximately 95% of flights have a minimum N1 % below 30%.

The upper bound for maximum N1 % application during braking is 90% while a few flights have maximum N1 % application below 40%. For these few flights, it is assumed that thrust reversers were not applied.

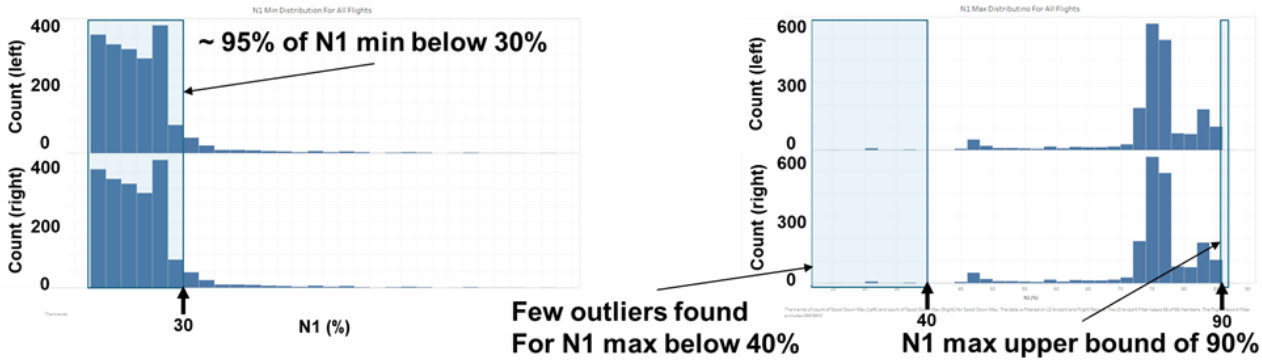


Figure 21. N1 % distribution during braking rollout for the NB-B2/2 airframe with CFM engines

To model the spool-down time in terms of N1 %, the timestamps of N1 % at 10% interval are subtracted from the time stamp of N1 at 30 %. To analyze the complete spool-down time, the model is trained to represent spool-down time for N1% between 90% and 30%. There are two ways to collect the spool-down time from N1 % of 30%. One way is to collect time for N1 % from 80 % to ~ 30 % in 10 % increments. This means finding the time for N1% from 80 % to 30 % with a small margin (+/- 1%) at the 80 % mark, then finding the time for N1 % from 80 % to 30 % with the same margin at the 80 % mark, and so on. The other way is to collect the expected time of N1 % distribution at 10 % interval at 10 % increments. This means finding the expectation of the distribution of time from N1 % between 80% ~ 89% to 30 %, then finding the expectation of the distribution of time from N1 % between 70% ~ 79% to 30 %. The first data collection method would result in a smaller number of data to train the spool-down time model since spool-down time data is not collected from flights that do not reach a specific N1 % threshold. For example, some flights that deployed thrust reverser did not deploy N1 % up to 80%. However, this spool-down time represents the time to a specific N1 % value so the model may be more accurate. The second data collection method would result in a larger number of data to train the spool-down time model since the expectation of the spool-down time distribution is found when reaching any point in the N1 % intervals. However, there may be some error in the spool-down time model based on the N1 % that each interval may represent.

5.2 Results

The first data collection method is used to generate spool-down time buckets that correspond to N1 % going from 80% to 30% in 10% increments (first time corresponds for N1 % from 80% to 70%, second time corresponds to N1 % from 80% to 60%, etc.). Average and median values of the time difference for reaching N1 % of 30% from N1 % of 40 % to 80 % thresholds are used to create a regression model on spool-down time. This model and the data distribution are illustrated in Figure 22.

Upper N1 margin (%)	Median (s)	Average (s)	stddev	Min	Max
40	3	2.769	1.852	0	32
50	4	4.618	3.951	0	46
60	5	5.877	4.558	2	48
70	6	6.36	3.489	0	48
80	7	7.899	3.905	6	50

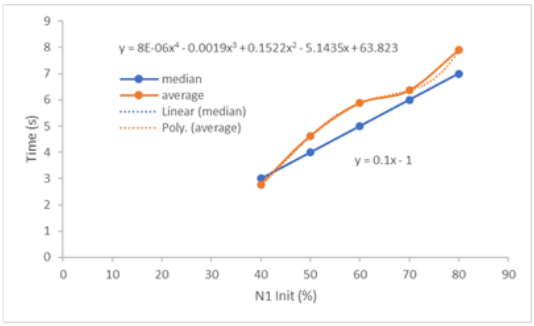


Figure 22. Median and average time to spool-down from N1 % of 30% to various levels of N1 % at 10% increments

A linear relationship between the median of reaching each N1 % trigger for each flight is observed. This linear model is defined by $T = 0.1\,N1 - 1$ where T is the spool-down time and $N1$ is the N1 %. A nonlinear relationship between average time to reaching each N1 % trigger for all the flights is observed. For the average model, an S-shaped relationship is observed where there is a steeper increase in time to spool-down for reaching N1 of 40 % to 60 %. The relationship flattens out between 60 % to 70 % and then steepens again between 70 % and 80 %. When this relationship is modeled with a linear model, a parallel line is observed between the average and the median model with a small offset. This indicates that overall, a linear relationship can summarize the spool-down time model.

The second data collection method is used to generate the spool-down time model. The spool-down time model is generated based on the expectation of the distribution of time for each N1 % interval. The expectation can be found with the median or the average of the distribution. The expectation of time to N1 % at each N1 % interval can be connected to model the spool-down time. This model and the data distribution are illustrated in Figure 23.

Upper Window (%)	Median (s)	Average (s)	stddev	Min	Max
40-50	3	4.24675	4.21824	0	28
50-60	4	5.20833	3.42679	2	14
60-70	5	6.4	3.66465	4	24
70-80	6	6.13524	2.51477	2	46
80-90	7	7.83281	3.72561	6	50

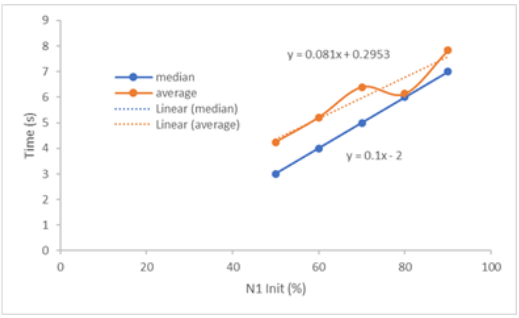


Figure 23. Median or average time to spool-down until N1 % reaches the expectation of N1 % interval, in 10 % increments

Again, a linear relationship between median and each N1 % window is observed with the median data. This linear model is defined by T = 0.1 N1 – 2 where T is the spool-down time and N1 is the N1 %. A nonlinear relationship between average time to N1 % and N1 % is observed. However, a linear regression line, T = 0.081 N1 – 0.295, does fit the average time to N1 % model as well. This indicates that overall, there is a linear relationship between N1 % and time to reach that N1 %.

6 Unsupervised learning

6.1 Identification and implementation of metrics

Prior to diving into the details, it is important to define some pertinent terms that are used frequently in order to avoid ambiguity. The following definitions from ASTM International Standard Terminology for Aircraft Braking Performance are used in this research.

1. **Aircraft braking coefficient**: the ratio of the deceleration force from the braked and unbraked wheels of a braked aircraft relative to the sum of the vertical (normal) force acting on the aircraft. The aircraft braking coefficient is determined by using the weight of the aircraft (W-L) and encompasses all the braking forces of all the gear, even those that are not braked.
2. **Wheel braking coefficient**: the ratio of the deceleration force from the braked wheels/tires relative to the sum of the vertical (normal) forces acting on the braked wheels/tires. The wheel braking coefficient is the result of the combination of all functioning braked wheels.
3. **Braking action**: a means of describing the maximum capability of a vehicle braking system on a wet or contaminated surface that references a standardized reporting scale.

4. **Pilot braking action report (PIREP), Aircraft Reports (AIREP)**: a report describing a level of braking action resulting from the observations of a pilot.
5. **Airport friction measurements**: the value obtained through ground measurement devices approved for use in measuring runway surface friction characteristics.

Flight parameters from the recorded data are used to define metrics of interest for assessing the braking performance of aircraft on contaminated runways. These metrics may be divided into three categories:

1. **Single point metrics**: calculated directly from fused data at a single point; might need metadata information to find that point; used in more complex metrics
2. **Metadata metrics**: calculated by extracting the metadata information and then encoding it as categorical information
3. **Calculated metrics** (includes time series measurements if any): calculated using fused flight data and metadata to obtain the required value(s); potentially using single point metrics

Table 4 summarizes single-point metrics and their descriptions. Table 5 summarizes metrics based on metadata and their descriptions. Table 6 summarizes metrics calculated from FOQA flight parameters and metadata, and their descriptions.

Table 4. Single point metrics

No.	Metric	Description
At Touchdown		
1	Ground Speed	Ground speed of the aircraft at touchdown (kts)
2	True Airspeed	Airspeed of the aircraft at touchdown (kts)
3	Vertical Speed	Vertical speed of the aircraft at touchdown (ft/s)
4	Weight	Weight of the aircraft at touchdown (lbs.)
5	Fuel Weight	Weight of the fuel left in the aircraft at touchdown (lbs.)
6	Specific Kinetic Energy	Specific kinetic energy of the aircraft at touchdown (lbs ft^2)
7	Number of g's	Vertical acceleration experienced by the aircraft at touchdown
8	Flap Setting	Angle of the flaps used at touchdown (degrees)
9	Headwind/tailwind	Wind experienced by the aircraft at touchdown (kts)
During Rollout on Runway		
10	Ground Speed at Runway Exit	Ground speed of the aircraft as it exits the runway after landing (kts)

11	Ground Spoilers Armed	Use of spoilers on the ground after landing
12	Autobrakes Armed	Use of autobrakes after landing
13	Thrust Reversers Deployed	Use of thrust reversers after landing
14	Thrust Reversers Stowed	Stop use of thrust reversers after landing
15	Thrust Reversers Peak Usage	Maximum engine rotation rate when thrust reversers are in use after landing

Table 5. Metadata Metrics

No.	Metric	Description
Airport Information		
1	Airport ID	Airport identification and information
2	Landing Airport ID	Identification of the information about the airport where the aircraft lands
3	Outside Air Temperature	Outside air temperature at the airport considered
4	Elevation	Elevation of the airport considered
5	Visibility	Visibility at the airport considered
Runway Information		
6	Runway ID	Runway identification and information
7	Runway Type	Treatment of the runway on which the aircraft lands (can be porous friction course (PFC), grooved, non-grooved, etc.)
8	Runway Condition Code	Condition code of the runway on which the aircraft lands
9	Runway Slope	Longitudinal slope of the runway on which the aircraft lands
Aircraft Information		
10	Airframe ID	Airframe identification and information

Table 6. Calculated Metrics

No.	Metric and Description
Approach	
1	Average Crab Angle during Approach (degrees)
2	Approach Stability
3	Altitude when Autopilot is Disconnected (ft)
Touchdown	
4	Speed Bleedoff Between 500 ft Above Ground and Runway Threshold (kts)

5	Height at Runway Threshold (Actual, Operational) (ft)
6	Speed Bleedoff Between Runway Threshold and Touchdown Point (kts)
7	Difference Between Published Touchdown Point and Height at Runway Threshold (ft)
8	Distance Between Runway Threshold and Touchdown Point (ft)
9	Seconds to Reach 40kts after touchdown (s)
10	Distance of Runway Used for Flare (ft)
Braking	
11	Average Aircraft Deceleration During Rollout (ft/s^2)
12	Standard Deviation of Deceleration During Rollout (ft/s^2)
13	Thrust Reverser Usage During Rollout
14	Time to Reach Maximum Reverse Thrust after Touchdown Point (s)
15	Ground Spoiler Deployment During Rollout
16	Seconds to Brake Use (Manual or Autobraking) (s)
17	Distance to Runway Used for Braking (ft)
18	Average Brake Pressure During Rollout (psi)
19	Standard Deviation of Brake Pressure During Rollout (psi)
20	Brake Pressure Peak (using 2 seconds before and after peak) (psi)
21	Location of Runway Exit Compared to Runway Length
22	Main Gear Normal Force (time series from touchdown to runway exit) (lbs. ft^2)

The metrics identified in Table 4, Table 5, and Table 6 can be evaluated for each individual flight operation in the dataset. Figure 24 provides a notional representation of how this process works.

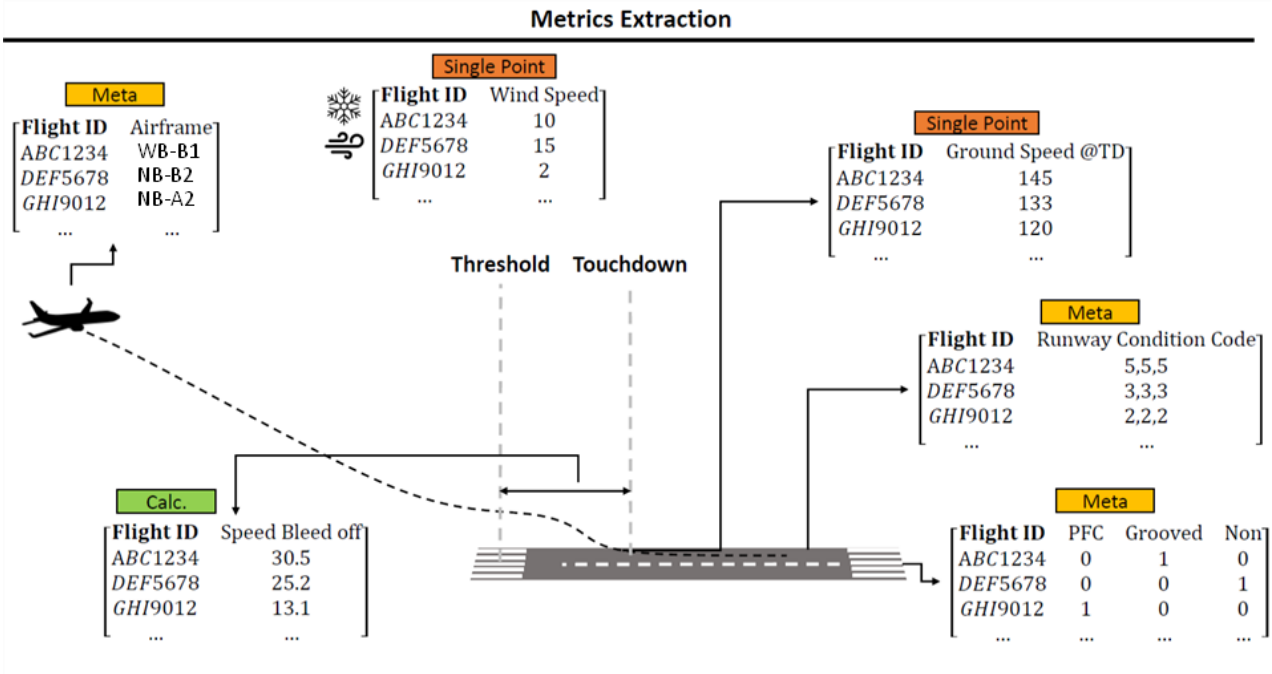


Figure 24. Notional representation of the extraction of metrics from routine flight operations

For single point and calculated metrics, the appropriate segments of the approach and landing phases are identified within the flight data recording. The metric is then calculated using the data recorded from each individual flight. The metadata metrics are constant for a particular flight operation independent of the data collected during the flight. The metrics collected for each flight record can be collated into a large high-dimensional feature vector as shown in Figure 25.

	Meta					Single Point	Calc.	
Flight ID	**Airframe**	**PFC**	**Grooved**	**Non**	**$RwCC_1$**	**Ground Speed TD**	**Speed Bleed Off**	...
*ABC*1234	WB-B1	0	1	0	5	145	30.5	...
*DEF*5678	NB-B2	0	0	1	3	133	25.2	...
*GHI*9012	NB-A2	1	0	0	2	120	13.1	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...

Figure 25. Notional depiction of feature vector matrix and its assembly

Each row in Figure 25 represents a unique flight operation that has been characterized by its extracted metrics rather than other parameters. All the flight operations in the datasets available may thus be collected together as a large matrix containing the values of the metrics as columns.

This allows for an easy comparison of the values of different metrics across flight operations and enables the visualizations that were shown earlier in Figure 13 and Figure 14. Additionally, this feature vector matrix is also the basis for building machine learning models using the data which will be the subject of future research. Collection and organization of the data in such a manner allows for data from flights operating at different airports, in different weather conditions, with different airframes, etc. to be easily compared against each other and existing standards.

6.2 Benchmark of clustering algorithms

Analyzing braking behavior during ground roll is important for evaluating runway safety. Conventional methods of analyzing braking behavior include measuring braking coefficient experimentally or estimating braking coefficient based on the physics behind the braking process. Both of these methods provide a general evaluation of the braking performance which can be used to find the performance limits or norm. However, they do not provide an analysis of braking performance sensitivity to flight parameters and runway condition deviations from the idealized aircraft braking scenario. This problem can be resolved by characterizing the braking behavior with machine learning from real flight data. This is made possible by the explosion of flight data available for analysis from an increased number of flights and improved computational power onboard the aircraft.

Machine learning algorithms can be categorized as *unsupervised* or *supervised* algorithms. Supervised machine learning relies on truth data to train a model to predict or explain behavior. The benefit is that supervised algorithms can provide a targeted analysis of the features of interest. However, the truth data may be difficult to obtain. Unsupervised machine learning utilizes the input data to create a model that formulates assumptions about the data. These assumptions may be related to the distance, sequence, or any noticeable feature recognized by the algorithm. Because truth data is not provided during the model training process, the analysis may or may not be relevant to the features of interest. However, the unsupervised algorithm can provide foundational knowledge about the data at hand. One such unsupervised learning method is able to cluster flights based on their similarities to derive knowledge about potential anomalies in the data.

The similarity of two datasets may be defined based on distance, density, or probability of the data distribution between samples. *Symbolic dynamic filtering* is an example of a distance-based anomaly detection method that calculates the pairwise distance between points in two data samples through the K-nearest neighbor method. Considering that all the flights landed safely and the goal of this study is to identify outliers, a one-to-one comparison of all the flights may be

a suboptimal choice for the analysis. Probability-based anomaly detection methods find the probability of point *p* outside of dataset *S*. The probability of a dataset with large dimensional data may be mapped to a lower dimension with algorithms such as support vector machines. The position of point *p* from a pre-defined threshold in this alternate dimension will determine whether the data point is an anomaly.

An example of probability-based algorithms is the Multiple Kernel Anomaly Detection (MKAD) algorithm. Studies show that MKAD is more sensitive to discrete parameters. Since the available flight data contains both discrete and continuous parameters, probability-based algorithms may not be the best choice for analysis. Density-based anomaly detection algorithms such as the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) rely on spatial clustering of data at a lower dimension. The clusters are applied such that a minimum number of samples lie within proximity. Density-based algorithms are known to perform better with continuous parameters. Considering that a larger number of data features used to represent the braking performance of the aircraft are continuous, density-based algorithms are chosen for this analysis.

6.3 Clustering framework implementation and results

The goal of this study is to explore the calculated and single-point metrics of 11,620 flights using unsupervised machine learning. The first attempt to visualize this data at a surface level is through the T-distributed stochastic neighbor embedding (t-SNE) algorithm. The t-SNE algorithm maps the n-dimensional data to 2-dimensional space. The augmented dimension itself is meaningless. However, the relative distance of each data point in this dimension can be used to identify similar flights. The t-SNE algorithm does a surface-level analysis of the similarity between samples because it utilizes the input data features directly to augment the lower dimensions. Since t-SNE uses the input data features directly, the magnitude of values will contribute to the result. Hence, the input data is normalized by the min-max range of each feature. An illustration of the data flow is shown in Figure 26.

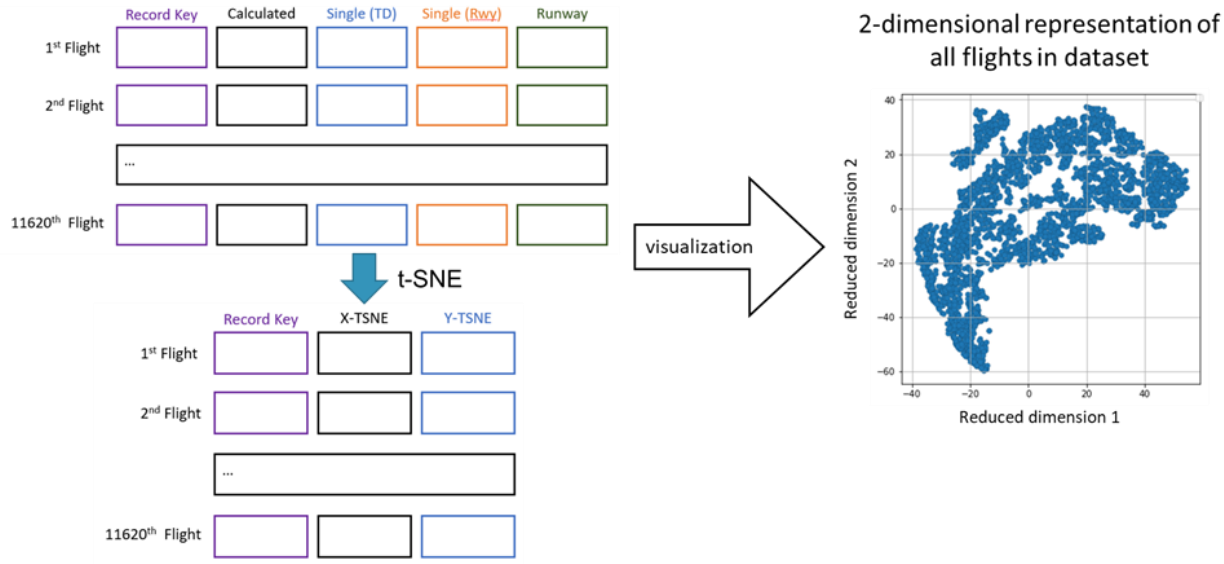


Figure 26. Illustration of the data flow for obtaining results from the t-SNE algorithm

A significant benefit of the t-SNE algorithm is that the lower dimensions generated can be mapped to the rectangular coordinate plane to visualize how the data gets clustered. And just like any other machine learning algorithm, the results from the t-SNE algorithm are dependent on the hyperparameter values. A grid search algorithm is used to find the optimal hyperparameter values based on its simplicity and success in optimizing various machine learning algorithms. The main hyperparameters optimized for the t-SNE algorithm are *perplexity* and *number of iterations*. A 3 by 3 search grid of perplexity and number of iterations is used to find the optimal configuration. The final values for these hyperparameters are found to be 100 and 1000, respectively. The final results are mapped with Boolean labels of various runway events to see if the clusters are formed based on known accident precursors. The runway events used in this analysis are *unstable approach*, *long landing*, *high energy descent* (at threshold), and *hard landing*. The events are flagged by definitions made by the data provider and are assumed to be true. The t-SNE results are also mapped with the airframe data to see if the clusters are airframe-specific.

An illustration of the t-SNE results mapped with the *long landing* precursor event and airframes is shown in Figure 27.

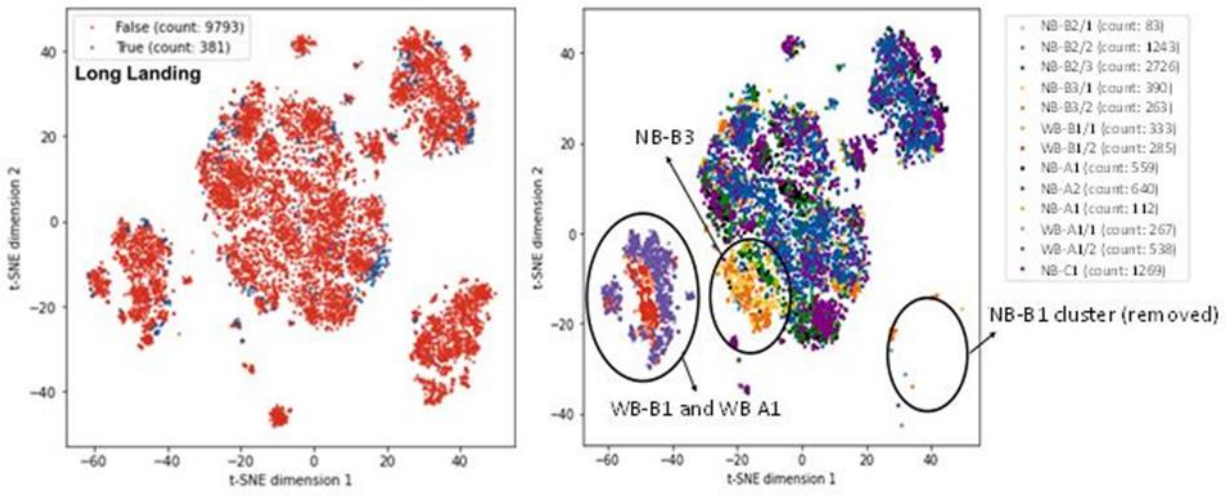


Figure 27. Results from unsupervised clustering t-SNE algorithm mapped with long landing precursor event and airframe type

Figure 27 shows that the t-SNE results are almost independent of the accident precursors since the precursors exist in all of the major clusters. This is expected since the features generated for unsupervised learning analysis are purposefully defined to characterize the braking performance. These features are not directly used for defining the precursors selected. However, the t-SNE results do have some relationship with the airframe type. This is expected since various airframes land and brake with specific system configurations. For example, the spoiler deflection during braking rollout is airframe-specific and the spoiler deflection is included in the input data. So even though the data features are normalized by the min-max range, there are many more features in the input data which are airframe-specific. This means the braking performance clustering analysis may need to be done per airframe so that airframe-specific feature variations do not dominate the clustering results. *Note:* On the right of Figure 27, the NB-B1 airframe data is removed to match the airframes used to generate the t-SNE clusters based on raw FOQA data.

Since density-based algorithms are selected for unsupervised machine learning, two algorithms, DBSCAN and Ordering Points To Identify the Clustering Structure (OPTICS) are selected based on their frequent use in anomaly detection. DBSCAN uses core distance and reachability distance to identify clusters. When a core data point is selected based on the relative proximity of nearby data points, the cluster is generated based on the predefined max range from the core point.

OPTICS hierarchically ranks the reachability of data points and a new cluster is defined at the point where the jump in reachability is observed. An illustration of the two algorithms is shown in Figure 28. The left-side of Figure 28 illustrates the process of finding the neighboring data points from the core data point based on a maximum distance threshold, denoted as epsilon (eps). The right-side of Figure 28 shows the hierarchical ranking of the data points based on proximity used to cluster similar flights. In this case, red, blue and green samples are clustered flights, and orange samples are outlier flights. Outliers are unique flights that do not show similar behavior against any other, leaving them unclustered. In terms of runway safety, outliers are flights that landed in an unusual manner (i.e., in a way not seen in other flights). Considering that aircraft are strongly encouraged to land in a specific configuration, outliers may represent flights that violate these guidelines and this indicates potential risk. Once the flights are clustered, the analysis focuses on determining the reason why clusters are formed and on generating a detailed breakdown of what makes the outliers unique.

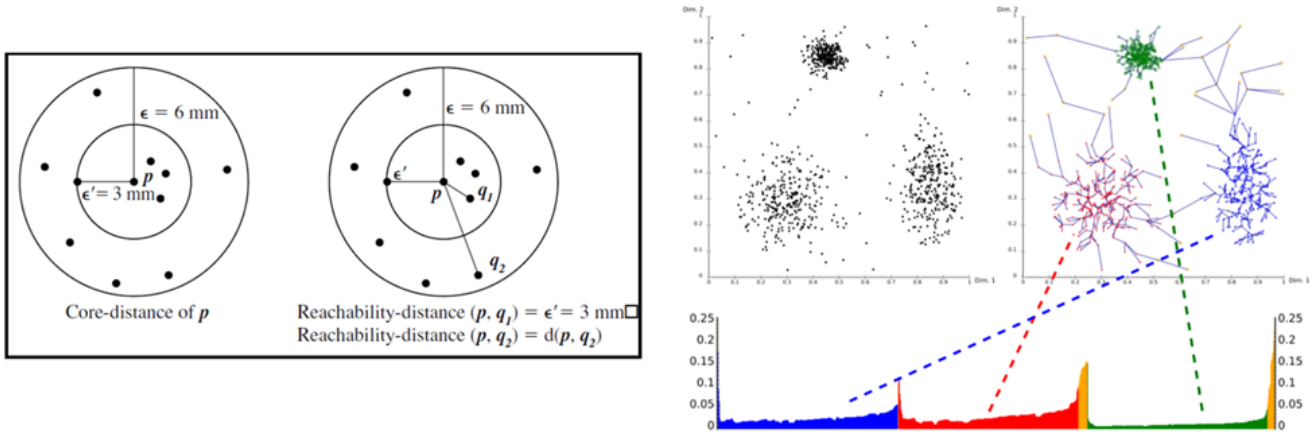


Figure 28. Illustration of DBSCAN and OPTICS algorithms and how clusters are generated

6.3.1 NB-B2 group

Since aircraft landing configuration varies significantly with airframe type, the clustering analysis is performed for each airframe group. The NB-B2 airframe group is selected because it represents the largest number of flights in the dataset. The data includes 83 NB-B2/1 variants, 1,243 NB-B2/2 variants, and 2,726 NB-B2/3 variants. The DBSCAN and OPTICS clustering algorithms are used to generate clusters and the hyperparameters are varied until the largest number of clusters are formed while maintaining the outlier percentage below 5%. This is because outliers represent unique landing configurations that potentially landed slightly differently from the other, nominal operations. However, it is known that all of the flights used in

this analysis landed safely. Therefore, it would not make sense to cluster them so that a larger portion of the flights are un-clustered. The DBSCAN and OPTICS clustering results for the NB-B2 airframe group are depicted in Figure 29.

The top graph shows the OPTICS results based on the hierarchical ordering of the data points in terms of proximity. The green and red dots are the clustered flights, while the black dots are the outliers. The bottom three graphs show the t-SNE results for the NB-B2 airframe flights mapped with OPTICS clusters (left), airframe type (middle), and DBSCAN clusters (right). As can be seen in Figure 29, DBSCAN and OPTICS generate very similar clusters because they use the same fundamental technique for clustering the flights. It is interesting to notice that the clusters are not formed based on the airframe type, so they need to be further analyzed to identify trends in the landing process. It should also be pointed out that the t-SNE results generate similar clusters as DBSCAN and OPTICS, thus indicating the potential for some features having a dominant contribution to the landing configuration.

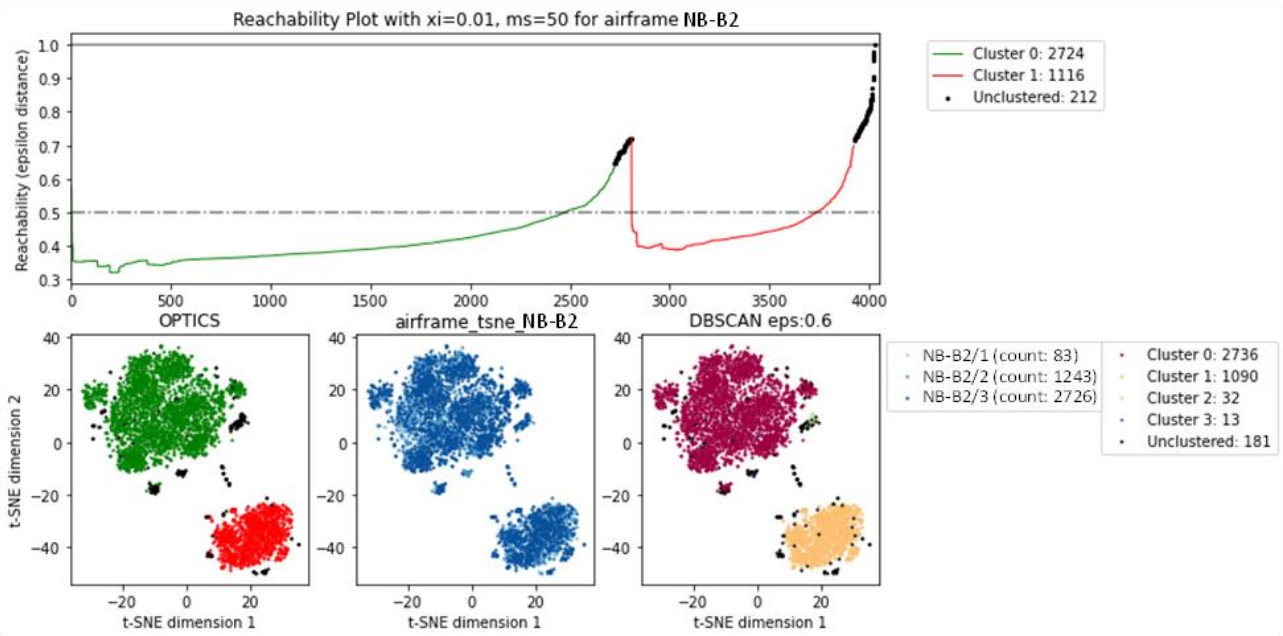


Figure 29. DBSCAN and OPTICS analysis results for flights of the NB-B2 airframe group

In order to determine why these specific clusters are formed, we need to look at the distribution of all the features used to generate the clusters. It can be observed that the clusters may be due to poor energy management. A significant deviation in true airspeed and ground speed at touchdown are also observed, as shown in Figure 30, potentially causing the smaller clusters 2 and 3 in Figure 29.

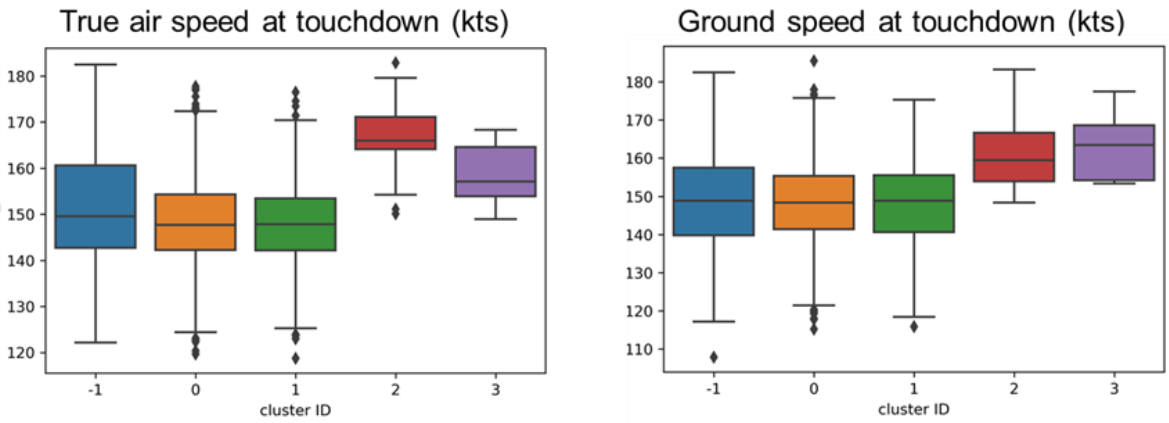


Figure 30. Distribution of true airspeed and ground speed at touchdown for flights of the NB-B2 airframe group

The difference between clusters 2 and 3 may be based on the level of wind experienced at touchdown, which indicates that poor management in the case of cluster 2 may be associated with unexpected wind. However, this alone does not have any implication on the fact that clusters 0 and 1 are generated, and no significant deviations in the features are observed. This may indicate an accumulation of small differences in multiple metrics between clusters 0 and 1 causing them to be differentiated. Further analysis focused on each airframe type may be required to differentiate the braking performance of each flight. Also, at this point, it is difficult to pinpoint why the outlier flights, denoted cluster -1, are un-clustered. Thus, an alternate analysis method is introduced in the next chapter to deep dive into each outlier.

6.3.2 NB-A1, NB-A2, NB-A3 group

A similar analysis is performed on the NB-A1, NB-A2, NB-A3 airframe group. The DBSCAN and OPTICS results are shown in Figure 31. This time, although the t-SNE analysis shows three distinct clusters of flights, both OPTICS and DBSCAN did not find significant differences from the two largest clusters and only found significant differences from the third small cluster. Again, the clusters are not associated with the airframe type which is beneficial considering that the focus of this research is on landing performance for any given airframe. The difference between the OPTICS and the DBSCAN algorithms is that DBSCAN treats the small third cluster as an actual cluster, whereas OPTICS treats it as a group of outliers. This is because the algorithm tries to find a jump in the reachability, and although the third cluster flights are further away from the others, they are not so far that they can be treated as separate from the main group. Instead, they are treated as outliers compared to the main cluster. It is interesting to point out that the third smaller cluster obtained from DBSCAN is divided into two even smaller clusters.

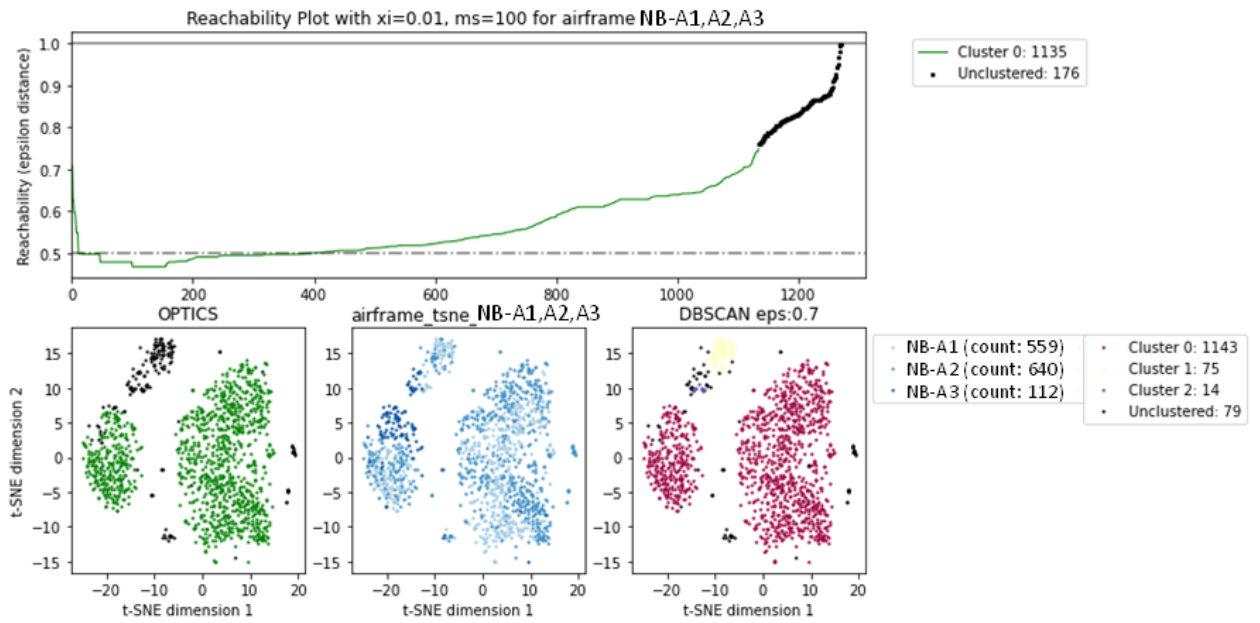


Figure 31. DBSCAN and OPTICS analysis results for flights of NB-A1, NB-A2, NB-A3 airframe group

To reduce the number of outliers to below 5%, the epsilon parameter is increased to 0.9. This results in an interesting behavior depicted in Figure 32. The new DBSCAN results show that a higher epsilon value, the smaller set of the third cluster seen in Figure 31 is treated as a separate cluster. The rest of the flights are grouped into a single larger nominal cluster. It is also worth noticing the scatter of the data points obtained from the t-SNE algorithm, for which the hyperparameters are kept constant. The shift in the scatter is due to the random number generator used by the t-SNE algorithm to generate the clusters: each time the algorithm is run, a slightly different scatter plot is generated. The question is then: “what makes these 12 flights so unique from the other flights.”

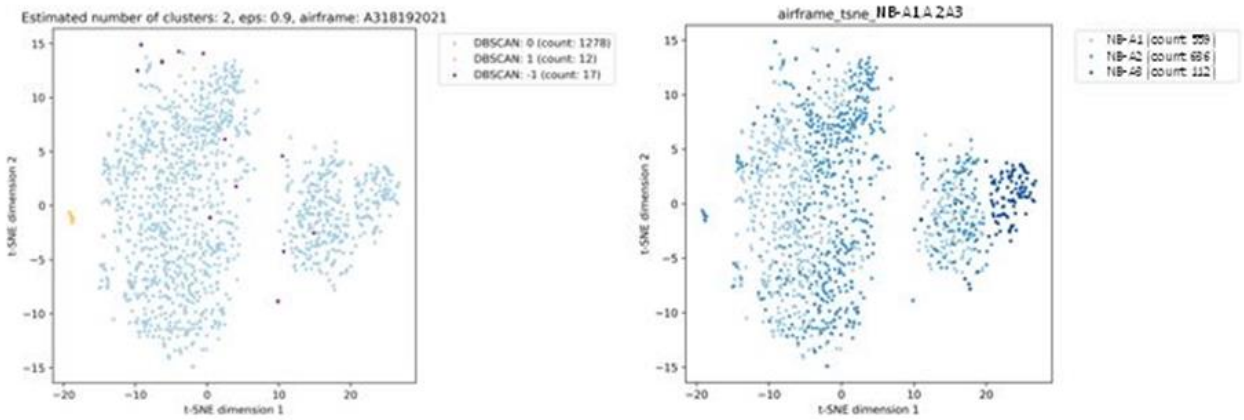


Figure 32. DBSCAN clusters generated with Epsilon = 0.9 for flights of the NB-A1, NB-A2, NB-A3 airframe group mapped to the t_SNE results

A quick look at the acceleration data distribution in Figure 33 shows a strange behavior in both lateral and normal average acceleration during rollout.

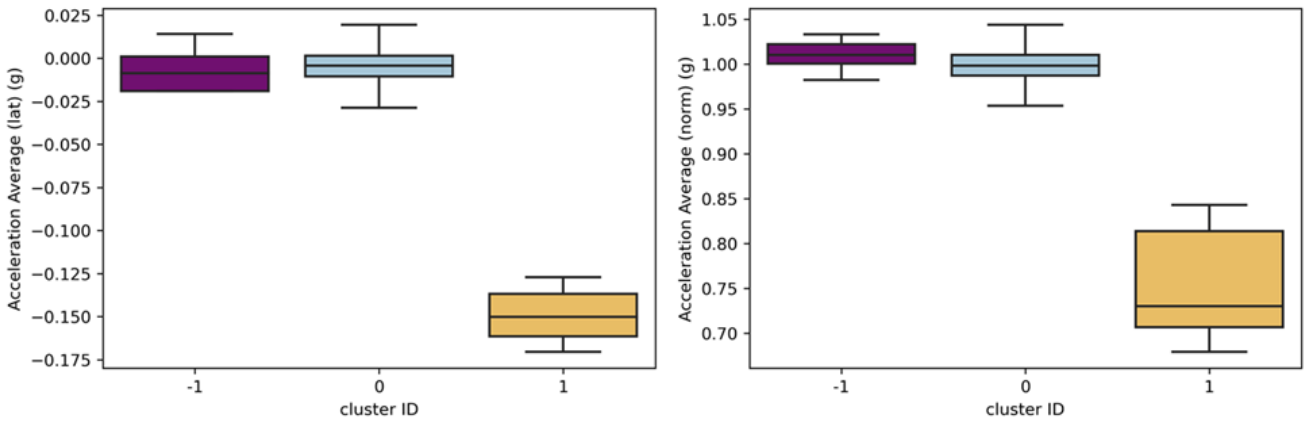


Figure 33. Average lateral and normal acceleration distribution during ground roll for flights of NB-A1, NB-A2, NB-A3 airframe group

The recording indicates a significant lateral motion which points to the potential of strong crosswind. However, there is a possibility of erroneous recording since, at a strong crosswind, the lateral deviation from the wind is corrected by a similar lateral deviation by deflecting the rudder. Another possibility is that these flights were assigned to an exit close to the runway threshold, so they only had a shorter portion of the runway to brake and exit.

Because the average acceleration is calculated between the touchdown point and the point when the aircraft speed reaches 40 kt, it is possible that in the scenario where an aircraft has to exit the runway early, it actually exits before slowing down below 40 kt. This theory is supported by the distribution of distance to threshold and percent ground track used to reach 40 kt from the touchdown point. Only a very small portion of the runway is then used to slow down, and these few flights need to touch down close to the runway threshold indicating that they may have to exit the runway early. This is observed in Figure 34.

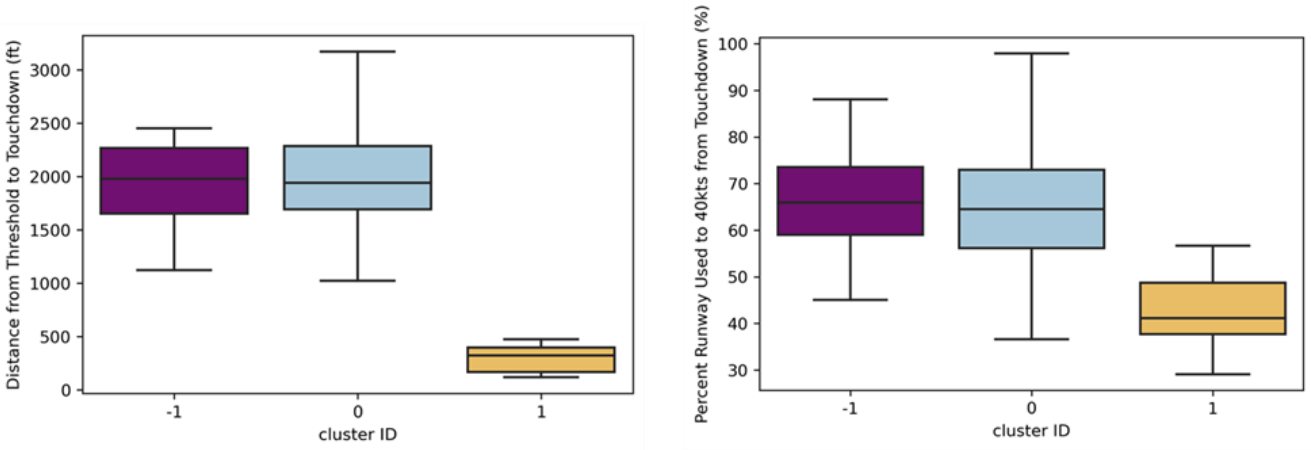


Figure 34. Distance from touchdown to threshold and percent runway used to slow down to 40 kt from touchdown for flights of the NB-A1, NB-A2, NB-A3 airframe group

Unlike the flights for the NB-B2 group, there are many indications as to why the outlier flights in the NB-A1, NB-A2, NB-A3 group are not clustered. Many control parameters vary significantly from the clustered flights as shown in Figure 35.

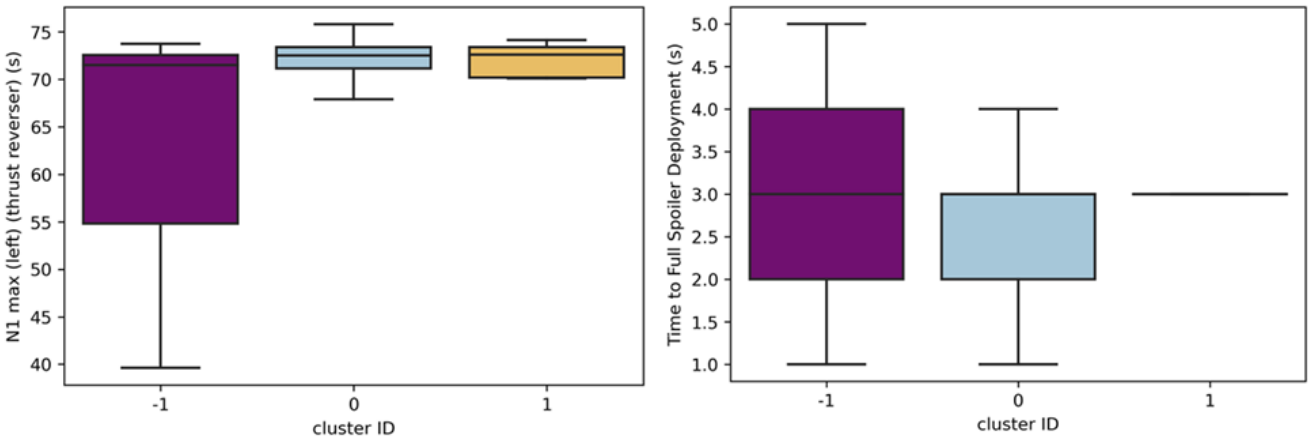


Figure 35. Max N1% and time to full spoiler deployment during ground roll for flights of the NB-A1, NB-A2, NB-A3 airframe group

The distribution indicates that the outlier flights may have not used thrust reverser or had delayed deployment of spoilers compared to the other flights. Since all the flights landed safely, the difference in the braking behavior must come from the piloting technique instead of from an error by the pilot. The percent runway used to slow down to 40 kt from the touchdown point metric seems to indicate this to be true, and the outlier flights used a similar amount of runway to stop as the nominal flights in cluster 0.

6.4 Observations and insights

An overall conclusion of the aforementioned analysis is that clustering the flights based on metrics that define the approach and landing performance may provide information on the overall landing pattern. These patterns are largely related to the decisions made during flight for energy management, available runway before exit, approach configuration, and more. However, a simple study based on unsupervised learning does not seem to be sufficient for reaching the goal of this study, which is identifying degraded braking with runway contamination. A few factors may be considered to improve the analysis.

First, we need to understand why flights are clustered in a specific way and which flight is associated with which cluster. Pilots need to follow a few general procedures for landing that may be slightly different for different airframe or airframe groups, so looking at metrics that summarize the overall landing behavior may not be enough for clustering algorithms to differentiate between flights with good or bad braking performance. For example, landing behaviors can include the level of control surface deflection, energy drained during final approach, and altitude at threshold crossing which may be associated with the pilots' expectation

of the braking performance, but not necessarily the braking experienced by the aircraft on specific segments of the runway.

Second, we need to evaluate the flights individually. Looking independently at each metric that significantly contributes to the way any given flight is clustered may point to the association of certain parameters. This kind of relationship is found for metrics that have larger deviation such as the level of control surface deflection and the amount of energy/speed drained during flare; or the percent runway used to slow down and the distance from the threshold to touchdown for early runway exit. Therefore, there is no doubt that clustering algorithms can differentiate flights, but the differentiation is sensitive to noise such as other metrics that are less relevant for the analysis. It is worth mentioning that this is why the unsupervised machine learning analysis is performed on each airframe group separately because initial studies showed that when all the flights are clustered simultaneously, the clusters are highly related to the airframe type.

7 Supervised learning Models

7.1 Decision trees and model interpretability

Decision trees classify a set of data based on a parameter threshold which is defined based on the "Gini" value. This value represents the cleanness of the classification (the lower the value, the better the split). This process is repeated until each leaf only contains data for a single cluster. It goes through multiple decision layers and at each layer it divides the data based on one data feature at one threshold. The divided data is passed to the next decision layer separately to go through the same process. A decision tree allows us to analyze high-dimensional data based on the interaction between the decisions made at each layer. These complicated interactions capture the few samples at the lower level, that would otherwise be difficult to categorize, which is representative of a typical flight event with multiple contributing factors. This benefit of the decision tree is used to pass the clusters obtained by the density clustering techniques to better interpret the clusters. An illustration of how the decision tree goes through multiple decisions to cluster flights is shown in Figure 36.

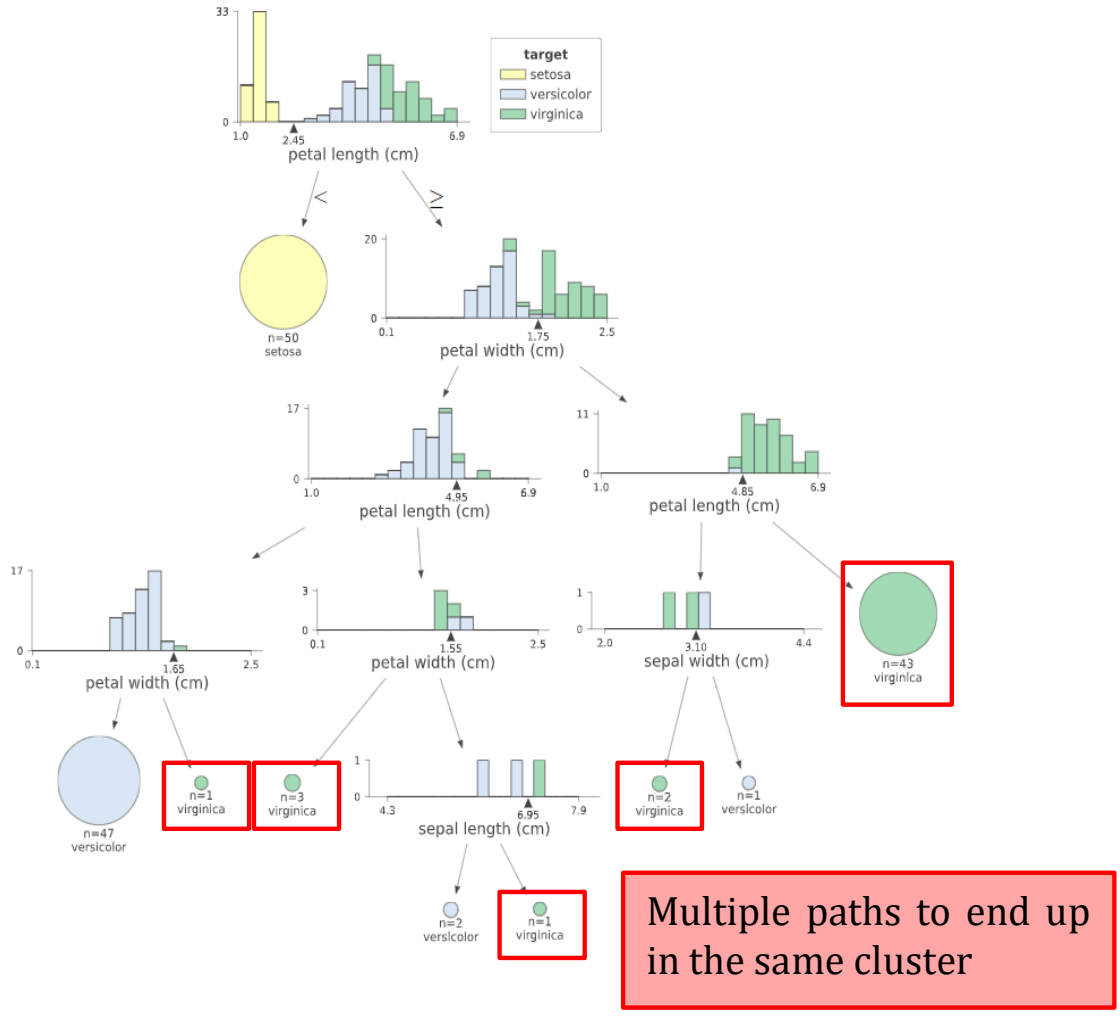


Figure 36. Illustration of decision tree clustering process

In the next section, we will showcase sample clustering results for the NB-B2/2 airframe and sample decision trees for the NB-B2/2 and NB-A2 airframes. Additional results for these and other airframes are provided in Appendix A: Remaining airframes clustering results and Appendix B: Remaining airframes decision tree results.

7.2 Decision tree for the NB-B2/2 flight data

Clusters generated by DBSCAN for the NB-B2/2 airframe group data show a need to further break down the flight data by each airframe type to analyze the anomalies in the landing process. DBSCAN is used to re-cluster the flights of the NB-B2/2 variant. Then, the decision tree method is used to analyze the clusters generated. DBSCAN results are depicted in Figure 37 and indicate

that most of the flights (1,189) are clustered into a single group. Only 21 flights are considered outliers based on their unique landing characteristics.

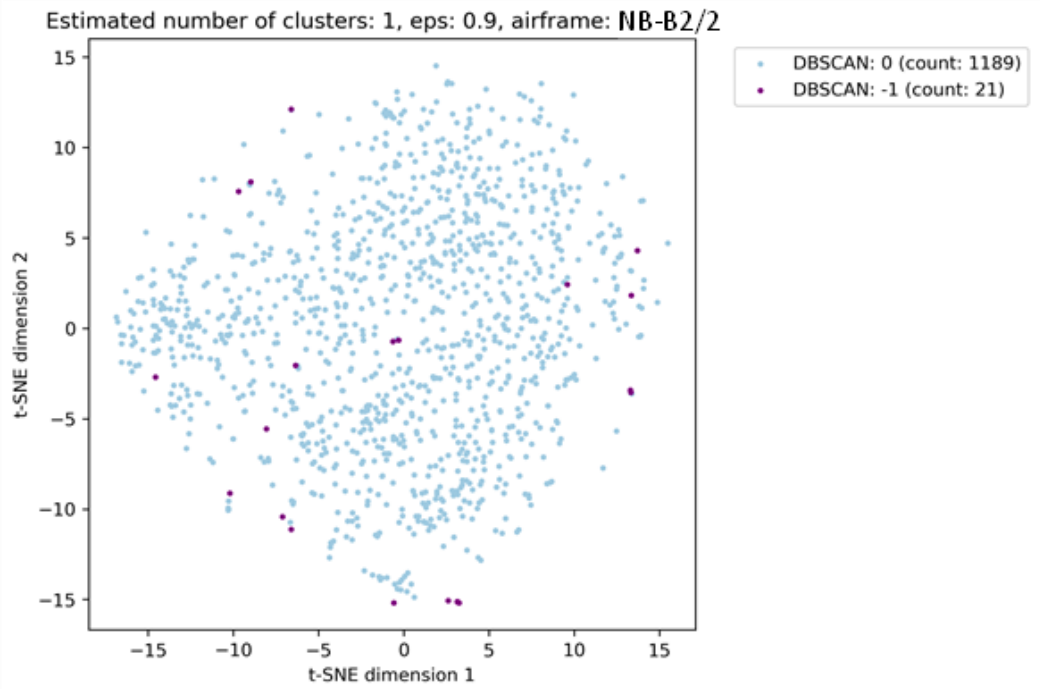


Figure 37. Illustration of the DBSCAN clusters generated for the NB-B2/2 airframe landing data mapped on the t-SNE results scatter plot

The cluster label data is used to train a decision tree classification model with the input data. The decision tree for the NB-B2/2 airframe is shown in Figure 38.

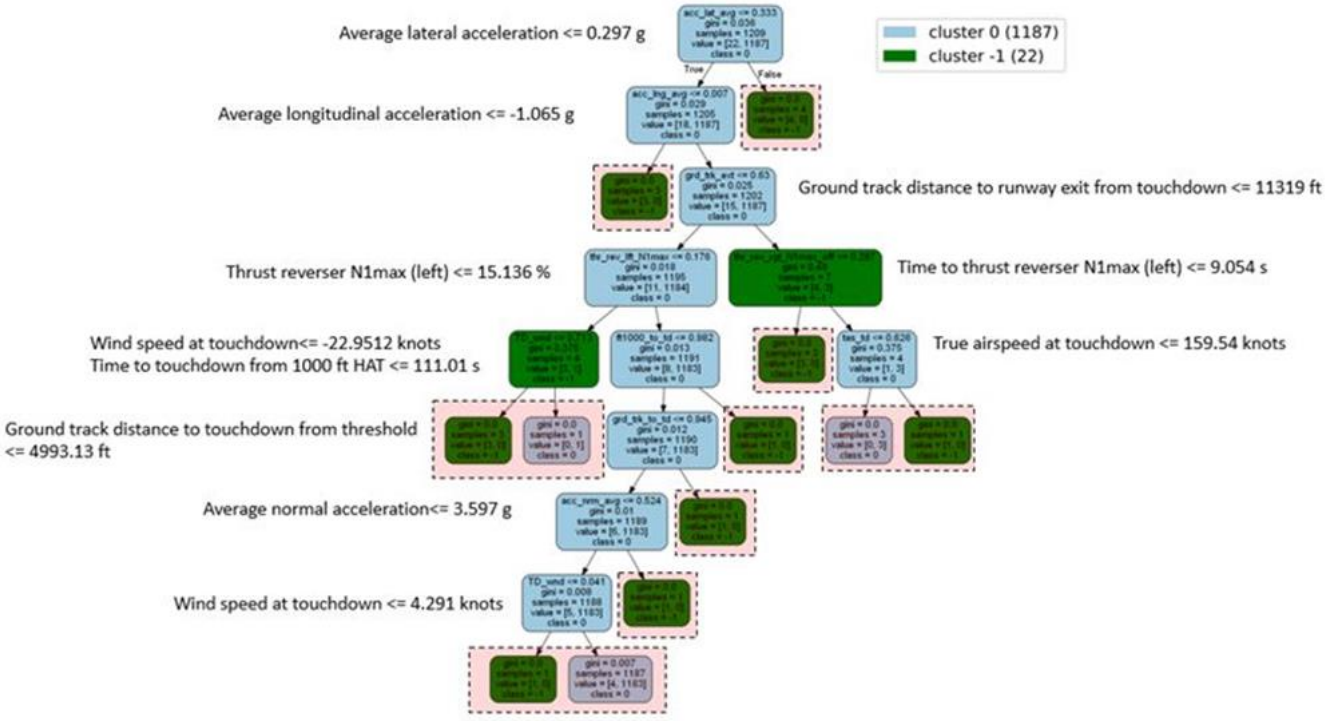


Figure 38. NB-B2/2 airframe landing data clusters decision tree

The immediate outliers are identified by recording errors in the acceleration data. Seven outliers identified at the top level and one outlier, all identified in Figure 39, have average longitudinal, lateral, and normal acceleration beyond physical possibility. This indicates an important characteristic where the tree is sensitive to false recordings.

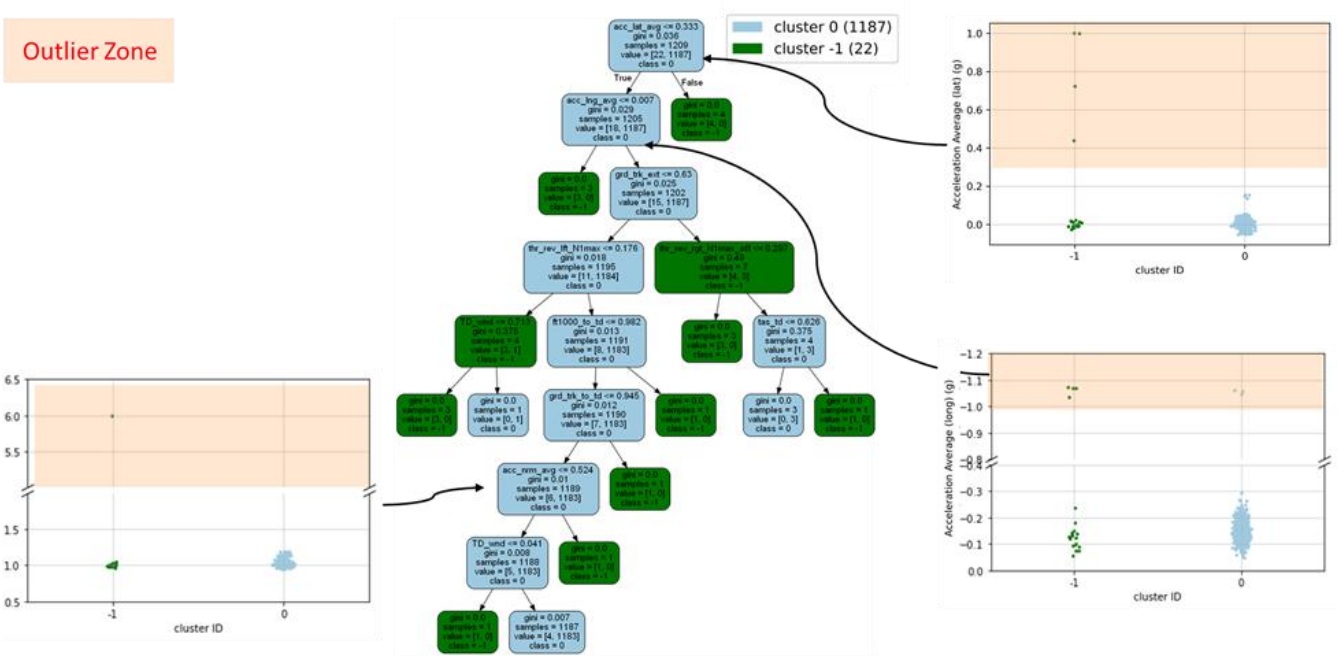


Figure 39. Outliers identified by the decision tree from erroneous recordings of acceleration for the NB-B2/2 airframe landing data

At the next level, the landing performance parameters are used to identify outliers. *Ground track distance to runway exit from touchdown* is an indication of how far on the runway the aircraft rolled before exiting the runway. It is also an indication of how long it took for the aircraft to slow down to 40 kt. However, a long rollout is not necessarily an indication of an anomalous behavior, but the decision tree identified that this combined with early application of thrust reverser is anomalous. The flights that did not take too long to brake or exit the runway from touchdown are further classified by the engine N1 %. Three outliers were marked with extremely low N1 % values (<= 15.14 %) indicating that these aircraft did not utilize thrust reversers at all, and the engine was kept at idle instead. At this point, the threshold values used by the decision tree come into question. The decision tree algorithm finds these values based on the resulting "Gini" value. However, when there are large gaps between two data sets that are clustered differently, any threshold values between these two clusters would result in a good "Gini" value. This means the exact threshold value used may not be meaningful. However, the sequence of decision parameters selected by the algorithm to cluster the data may introduce a coupling effect between parameters that are difficult for experts to identify without the aid of machine learning. So far, the decision tree in Figure 40 shows that the proper thrust reverser application is related to how long the aircraft needs to brake, which is an expected result. The first two decisions made

with regards to the acceleration parameters may be ignored since all they do is neglect extreme outliers from consideration.

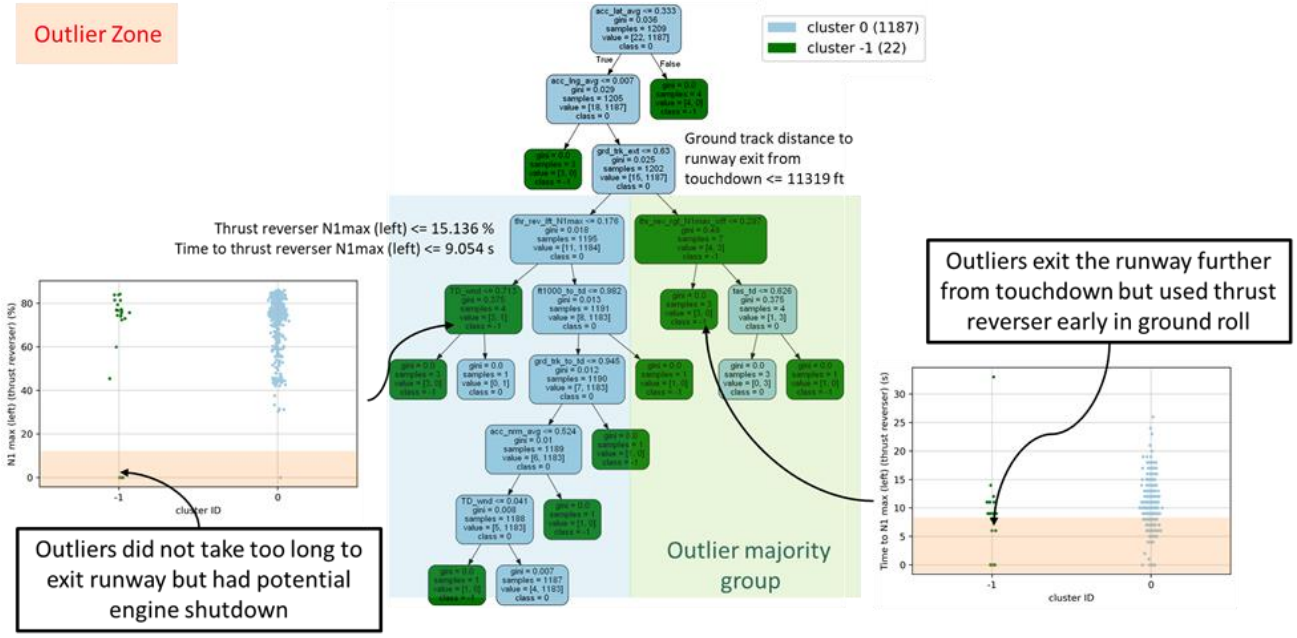


Figure 40. Outliers identified by the decision tree from proper application of thrust reversers and runway exit position for the NB-B2/2 airframe landing data

The next set of parameters that goes into identifying outliers in landing is related to wind speed and energy management, as depicted in the decision tree in Figure 41. At the lower levels, the wind speed parameters are used to find flights with high wind speed. It is worth noting that the direction of the wind speed is not important for the decision tree. Thus, as long as the headwind or tailwind is strong, the flights are tagged as outliers. Also, the decision tree generates clusters purely based on the data distribution instead of a reference. Therefore, 5 knots of wind may not be large in terms of the aircraft dynamics, but for the data set provided to the decision tree, this may be considered large. In terms of energy management, flights with high true airspeed at touchdown and long flare indicators are both used to flag outliers.

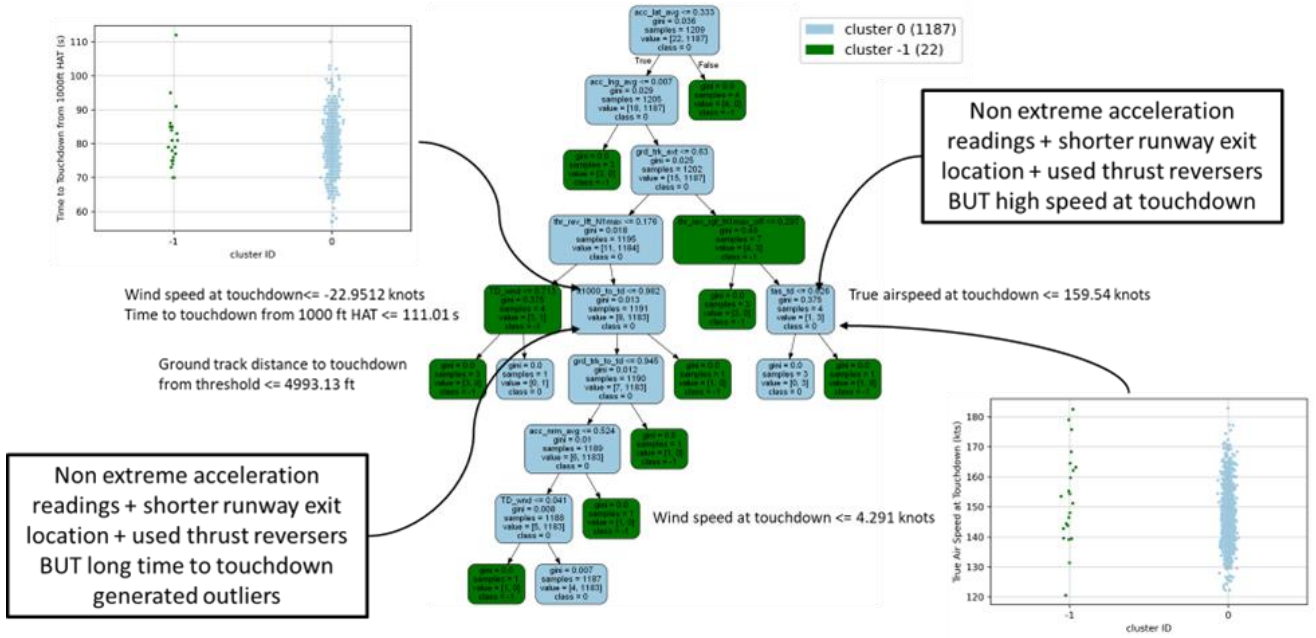


Figure 41. Outliers identified by the decision tree from wind speed, true airspeed, and time to touchdown from 1000 ft height above touchdown for the NB-B2/2 airframe landing data

Features of significance can be found from the decision tree based on the cleanness of the split from each feature. The significance order has some correlation to the level of depths that the decision parameter is used, as shown in Figure 42.

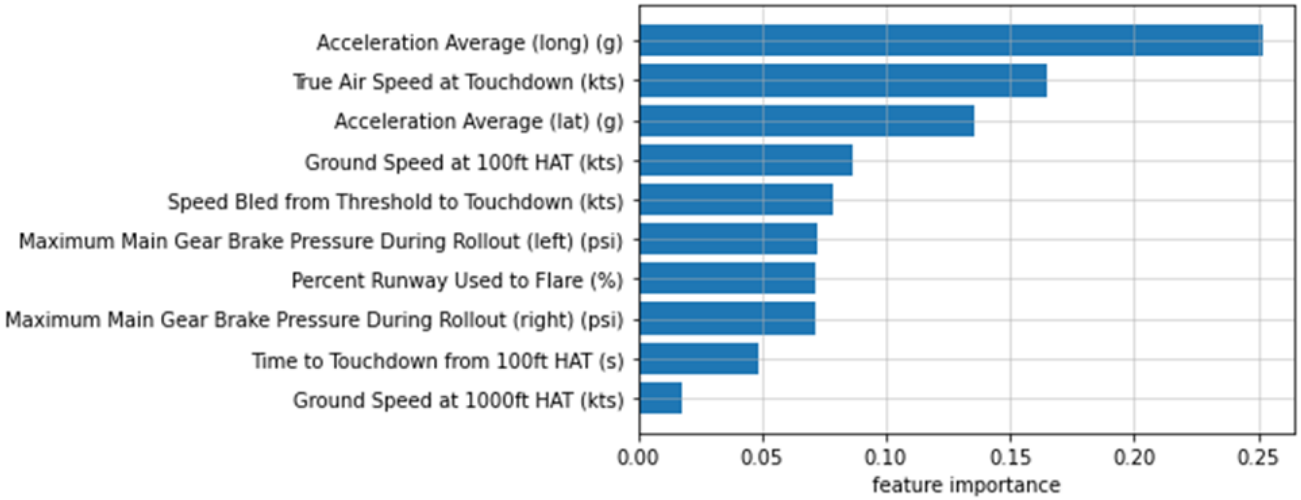


Figure 42. Feature significance values for the NB-B2/2 airframe data based on the cleanness of split from each decision parameter

Overall, the analysis from the first decision tree shows that the key takeaway is the potential metrics that contribute to the DBSCAN cluster generation. The coupling of these potential metrics may be considered for future flights to identify anomalies in terms of erroneous data recording, long landing, energy management, and proper braking device application.

7.3 Decision tree for NB-A2 flight data

The decision tree generated for the NB-B2/2 airframe data shows a deep tree with multiple couplings potentially impacting the landing performance. This may not always be the case as shown for the NB-A2 airframe flight data in Figure 43.

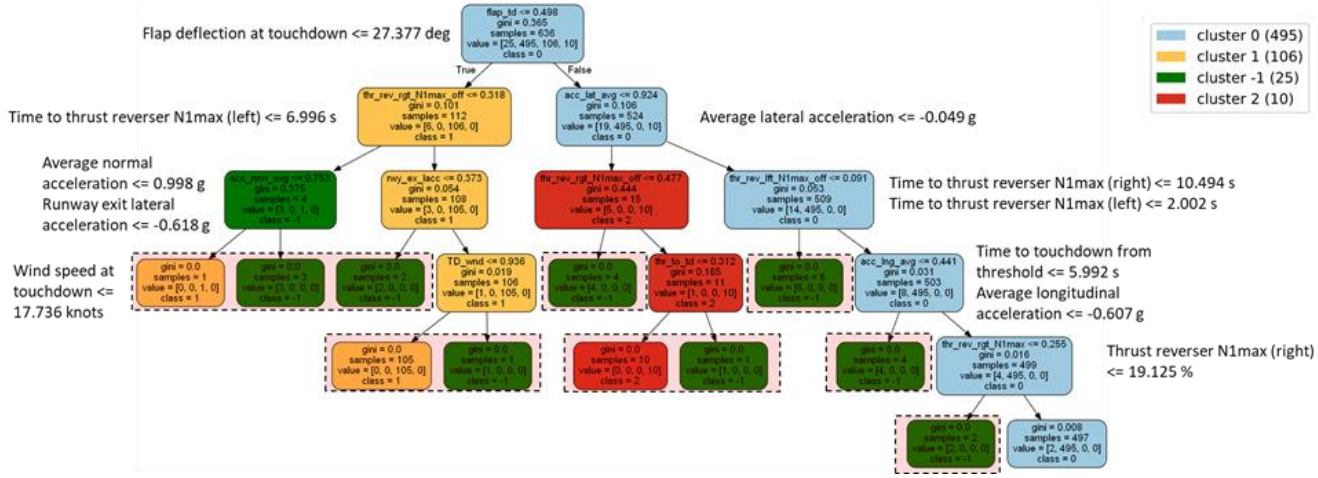


Figure 43. Decision tree for the NB-A2 airframe landing data clusters

The decision tree in Figure 43 shows a wide tree with multiple clusters. It is worth noting how each cluster typically lies within a single branch, while the outliers are leaves that fall out of each branch. This implies that outliers, in general, do not share a common characteristic. Thus, it is necessary to analyze each outlier or each small outlier group separately.

The first branch in light blue in Figure 43 represented by cluster 0 is flights that used full flap deflection and thrust reversers. This means that cluster 0 represents flights that followed the typical operating procedures and used all braking potential to slow down the aircraft. What makes cluster 1 flights unique from cluster 0 flights is their flap deflection. As shown in Figure 44, the flap deflection of cluster 1 flights is below 27.4 degrees which is lower than the recommended landing flap configuration. This is because some flights, when safe landing is expected, use lower flap deflection to save fuel. Besides this difference, cluster 1 also used thrust reversers, and had small acceleration and wind speed readings, which are indications of a typical safe landing. This shows that the clusters generated by DBSCAN do not only represent errors in

the recording and performance of the landing but also the pilot technique involved. Like cluster 1 flights, cluster 2 flights also show the ability of DBSCAN to identify variation in data due to pilot technique. More specifically, cluster 2 represents a difference in operation. The average acceleration is measured by acceleration recordings from touchdown to aircraft speed reaching 40 kt or exiting the runway.

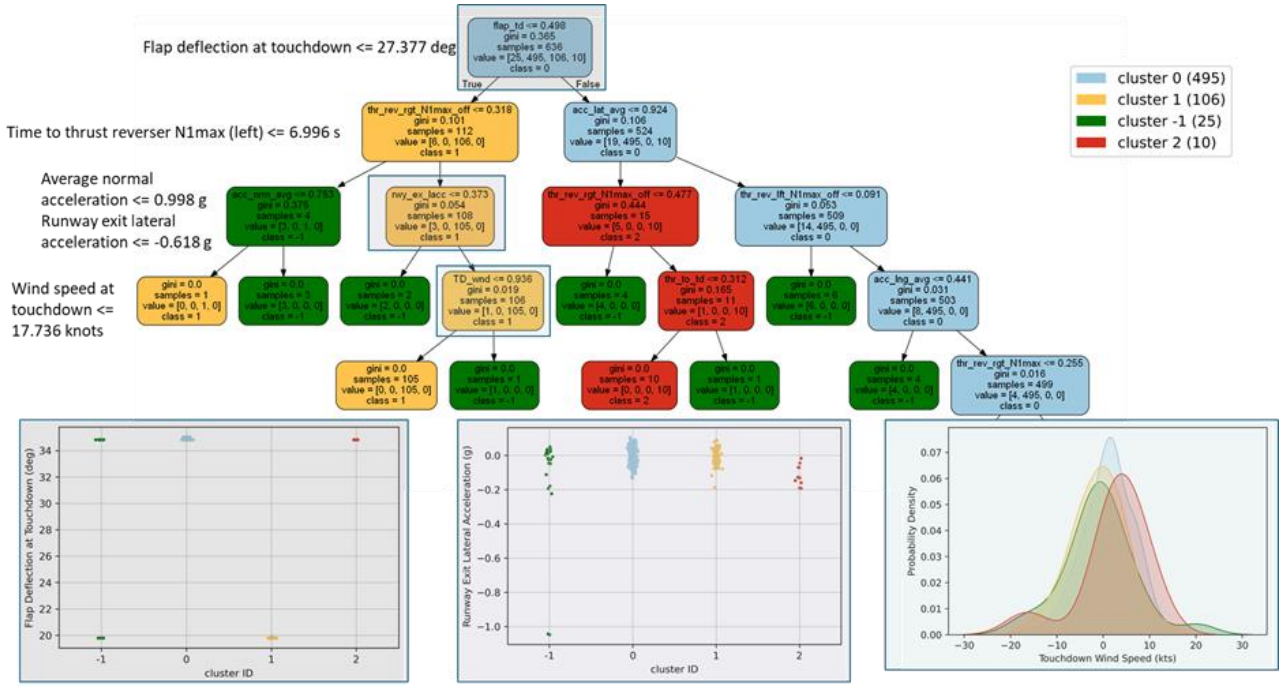


Figure 44. NB-A2 airframe flights DBSCAN cluster decision tree

With data distributions for flap deflection and runway exit lateral acceleration parameters, used to identify cluster 1 flights

An aircraft that exits the runway at the first few exits may exit the runway without reaching 40 kt. Also, the definition of "exit" is based on lateral deviation from the runway centerline beyond a threshold. When the two definitions combine, the large average lateral acceleration is due to the aircraft turning just before the exit. This is especially true when no significant lateral wind speed is recorded during ground roll. This is depicted in Figure 45.

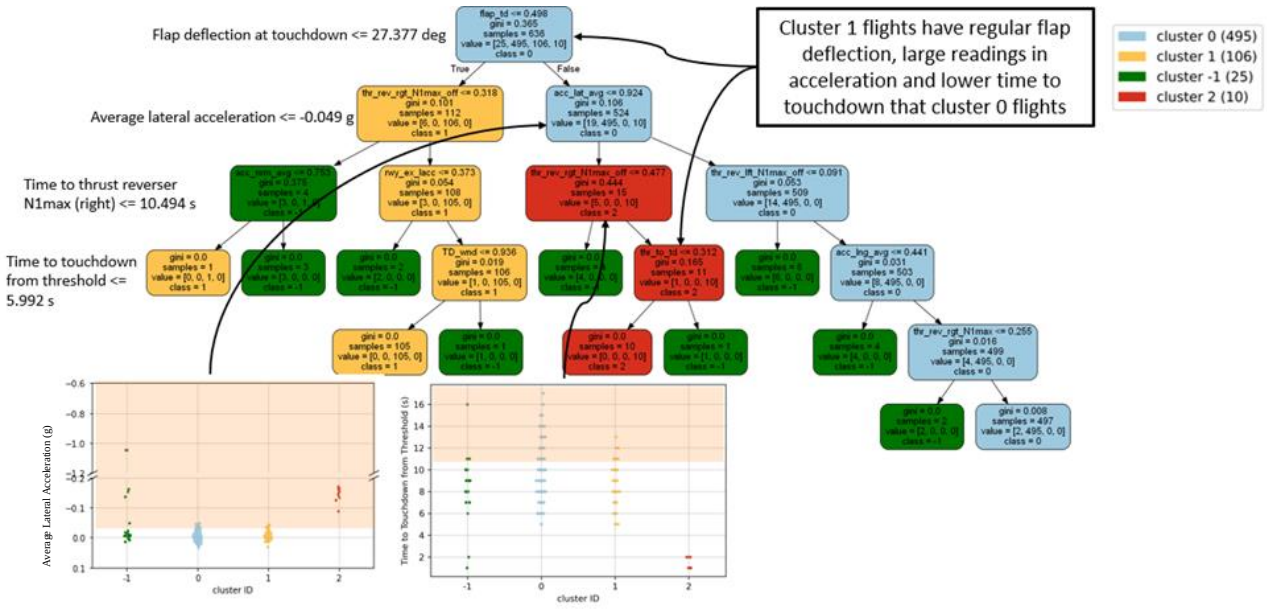


Figure 45. NB-A2 airframe flights DBSCAN cluster decision tree

With data distributions for average lateral acceleration and time to touchdown from threshold, used to identify cluster 2 flights

Based on this analysis, the outliers for the NB-A2 airframe flights are mainly determined by lack of braking device usage, large lateral acceleration when the aircraft exit is far from the threshold, and large wind speed. As depicted in Figure 46, another interesting trend is observed for cluster 2 flights. The height above touchdown metric shows a sudden jump just before the touchdown. This issue is found from the final decision layer for cluster 2 because the time to touchdown from threshold for these flights is less than a typical flight, as shown on the left of Figure 46. If the recording is accurate, then cluster 2 flights contacted the runway surface at high vertical touchdown speed followed by an immediate exit on the runway. This indicates a risk of operationally exiting the runway too early so there is a need to investigate the minimum distance to ground roll before runway exit.

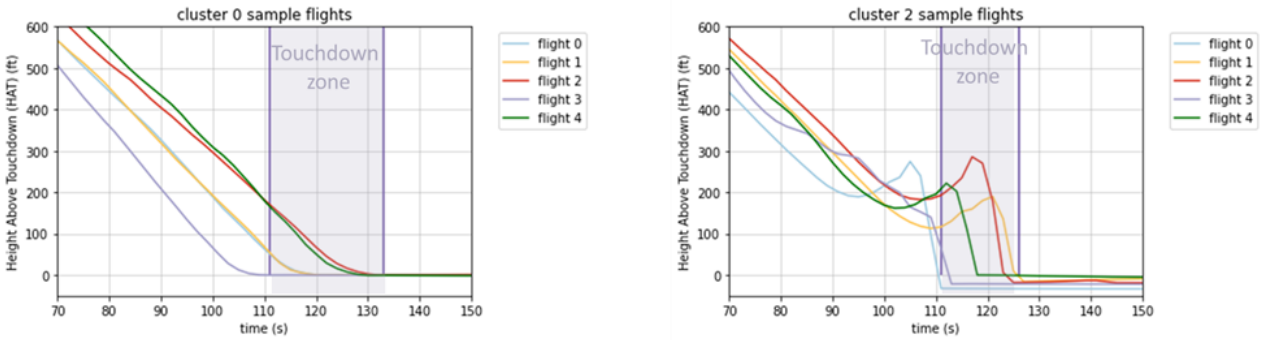


Figure 46. Sample height above touchdown trajectories for cluster 0 and 2 flights illustrating abnormal height above touchdown for cluster 2 flights

7.4 Frequently identified metrics and feature importance

The decision parameters and thresholds for various airframe flight clusters are shown in Table 7. This table indicates that 75% of all the calculated/single point metrics generated are used to determine the clusters according to the decision trees, while 30% of all the calculated/single point metrics generated are used by multiple airframes to generate the data clusters. Features of significance are calculated from the decision tree models based on the cleanness of splits generated by each decision parameter. The results show that 12 calculated and single-point metrics have a higher association to the landing performance and potentially degraded braking.

Table 7. Features and thresholds used by the decision tree to cluster the five airframe flight data based on DBSCAN results

Metrics (for outlier identification)	Threshold (NB-B2/2)	Threshold (NB-B2/3)	Threshold (NB-A2)	Threshold (WB-A1/2)	Threshold (NB-C1)
Acceleration Average (lat) (g)	0.002391	0.010466, 0.047797	-0.04624		
Acceleration Average (long) (g)	-0.0714, -0.27924, -0.10013				
Acceleration Average (norm) (g)		1.069162			1.362841
Average Crab Angle from 100 ft to Touchdown (rad)			6.852636		
Average Main Gear Brake Pressure During Rollout (left) (psi)				554.9545	
Distance to Runway Exit from Threshold (ft)		3668.262			
Flap Deflection at Touchdown (deg)		15.2	27.37669		
Fuel Weight at Touchdown (lbs)		22918.57			
Gross Weight at Touchdown (lbs)		158204.2		376397	
Ground Speed at 1000ft HAT (kts)	116.7169				
Ground Speed at 100ft HAT (kts)	150.6362				
Maximum Main Gear Brake Pressure During Rollout (left) (psi)	622.8029				
Maximum Main Gear Brake Pressure During Rollout (right) (psi)	2548.332, 1055.362				
N1 max (left) (thrust reverser) (%)			45.57288		54.9295
N1 max (right) (thrust reverser) (%)			49.9675		
Percent Ground Track Left at full Spoiler Deployment (%)					0.111869
Percent Ground Track Left at Main Gear Brake Application (%)		0.536372			
Percent Ground Track Left at N1 max (left) (%)		0.436886		0.887877	0.247257, 0.2219
Percent Ground Track Left at N1 max (right) (%)			0.704998	0.237211	0.664643
Percent Ground Track Left at Spoiler Deployment (%)		0.133636			
Percent Ground Track Used at Main Gear Brake Release (%)		0.988759			
Percent Runway Used to 40kts from Touchdown (%)			90.1318		
Percent Runway Used to Flare (%)	38.80042				7.345836
Runway Length (ft)			15253		
Speed Bled from Threshold to Touchdown (kts)	16.62217		8.943755		
Threshold Crossing Height (ft)					51.5, 52.5
Time to Touchdown from 100ft HAT (s)	16	19.502			
True Air Speed at Touchdown (kts)	178.1822, 175.6382		117.5545		
Vertical Speed at Touchdown (ft min)					2.690718

The second iteration of decision trees is generated by removing flights with poor recordings to observe if different clusters are generated. An example of this study is shown in Figure 47 for flights of the NB-B2/2 airframe. Table 8 indicates the decision parameters and thresholds used to identify the main cluster and the outlier flights. It is worth noting how acceleration metrics are still present but no longer used to identify poorly recorded flights. This is an important finding because some clustering information may be lost when fitting the decision tree model if certain parameters are used to make decisions for meaningless reasons. Besides this change, decision parameters similar to the ones used in the old model are observed. This indicates that similar decisions are still used to cluster outliers even after removing the poorly recorded flights. Based on the cleanness of the split made by each decision parameter, the decision tree is capable of evaluating the significance of each parameter.

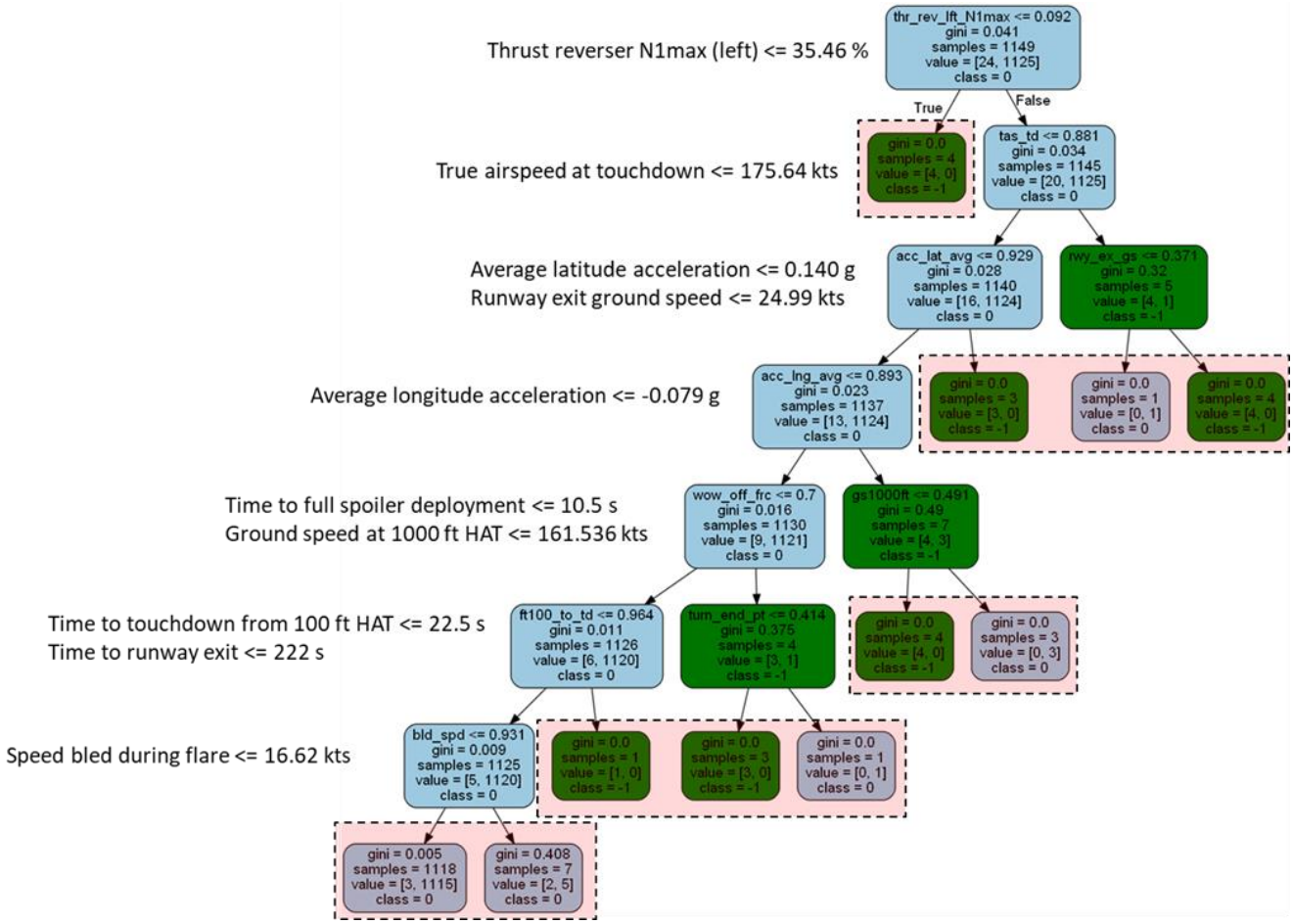


Figure 47. Second iteration of decision tree fitted with the NB-B2/2 airframe flights without bad acceleration recordings

Table 8. The NB-B2/2 decision parameters and thresholds for first and second iteration of decision trees

Metrics (for outlier identification)	Threshold (*with* extreme outliers)	Threshold (*without* extreme outliers)
Time to touchdown from 100 ft HAT (s)		<= 22.496
Ground speed at runway exit (kts)		> 24.989
Ground speed at 1000 ft HAT (kts)		<= 161.536
Speed bled during flare (kts)		16.6222
Time to runway exit (s)		<= 222.042
Average longitudinal acceleration during rollout (g)	<= 1.065	> -0.0795
Thrust reverser N1max (left) (%)	<= 15.136	<= 35.464
True airspeed at touchdown (kts)	> 159.54	> 175.638
Time to full spoiler deployment (s)		> 10.5
Average lateral acceleration during rollout (g)	> 0.297	> 0.140
Ground track distance to runway exit from touchdown (ft)	> 11319	
Time to thrust reverser N1max (left) (s)	<= 9.054	
Wind speed at touchdown (kts)	<= -22.951, > 4.291	
Time to touchdown from 1000 ft HAT (s)	> 111.01	
Average normal acceleration during rollout (g)	> 3.597	

8 Supervised Learning Applications

8.1 Aircraft braking performance and runway condition modeling

Factors such as weather condition, pavement texture characteristics, and runway slope all play critical roles in determining aircraft braking performance (OCallaghan, 2016). While past studies have explored how these factors may impact aircraft braking, they have not investigated the creation of quantitative models for evaluating braking action reports or runway friction characteristics. This section discusses using data fusion and statistical methods to quantitatively assess runway condition and aircraft braking action on contaminated runways.

The goal is to quantitatively assess pilot braking action reports (via Pilot Reports or PIREPs) and runway condition codes (RwyCCs). This is done by fusing multiple sources of data related to pilot braking actions, runway condition codes and characteristics, and prevailing weather conditions, and then building classification models using a machine learning approach.

Supervised machine learning techniques are effective to build classification models (Mair, et al., 2000; Hanyu, et al., 2018). According to the literature, random forest (Ali, Khan, Ahmad, &

Maqsood, 2012; Sekhar & Minal, 2016) and extreme gradient boosting (XGBoost) are identified as two supervised machine learning models that can achieve good classification results with a reasonably large training dataset and a set of fine-tuned hyperparameters (Chen & Guestrin, 2016; Zhang, et al., 2018; Chen, He, Benesty, & Khotilovich, 2019).

Supervised machine learning techniques are used to build classification models for two purposes:

1. To classify PIREPs given runway conditions and characteristics and weather conditions, denoted as Model #1 in this section.
2. To classify RwyCCs given pilot braking actions and runway characteristics, and weather conditions, denoted as Model #2 in this section.

8.1.1 Data description

The research presented in this section is based on the data source described in Section 3.1. However, since the supervised learning tasks in this section do not require the FOQA flight data, different approaches from the pipeline in Section 3.2. are used to obtain the actual input datasets used for the present tasks.

Two methods are used to fuse expanded FICON data shown in Figure 11 with ASOS data, resulting in two final fused datasets:

1. In the first case, expanded FICON is fused with concurrent weather data (weather corresponding to the FICON start time), known as “fused Dataset #1” in this report
2. In the second case, expanded FICON is fused with historical weather data (weather prior to the FICON start time, with a 1-hour time window), known as “fused Dataset #2” in this report.

In this second dataset, the fact that the ASOS data is recorded at 1-minute intervals results in 60 weather data points for each FICON record. The mean value of these points is fused with the expanded FICON. It is believed that the averaged ASOS data in the time window would make a reasonable representation of the weather condition prior to the FICON issuance and using this weather source can benefit the machine learning tasks afterwards, especially the RwyCC assessment task.

Figure 48 demonstrates the process of obtaining the two datasets. Figure 49 shows the format of their structures. Since they cover the main aspects of aircraft landing and braking procedure, the fused Datasets #1 and #2 are used as the foundation for building the classification models.

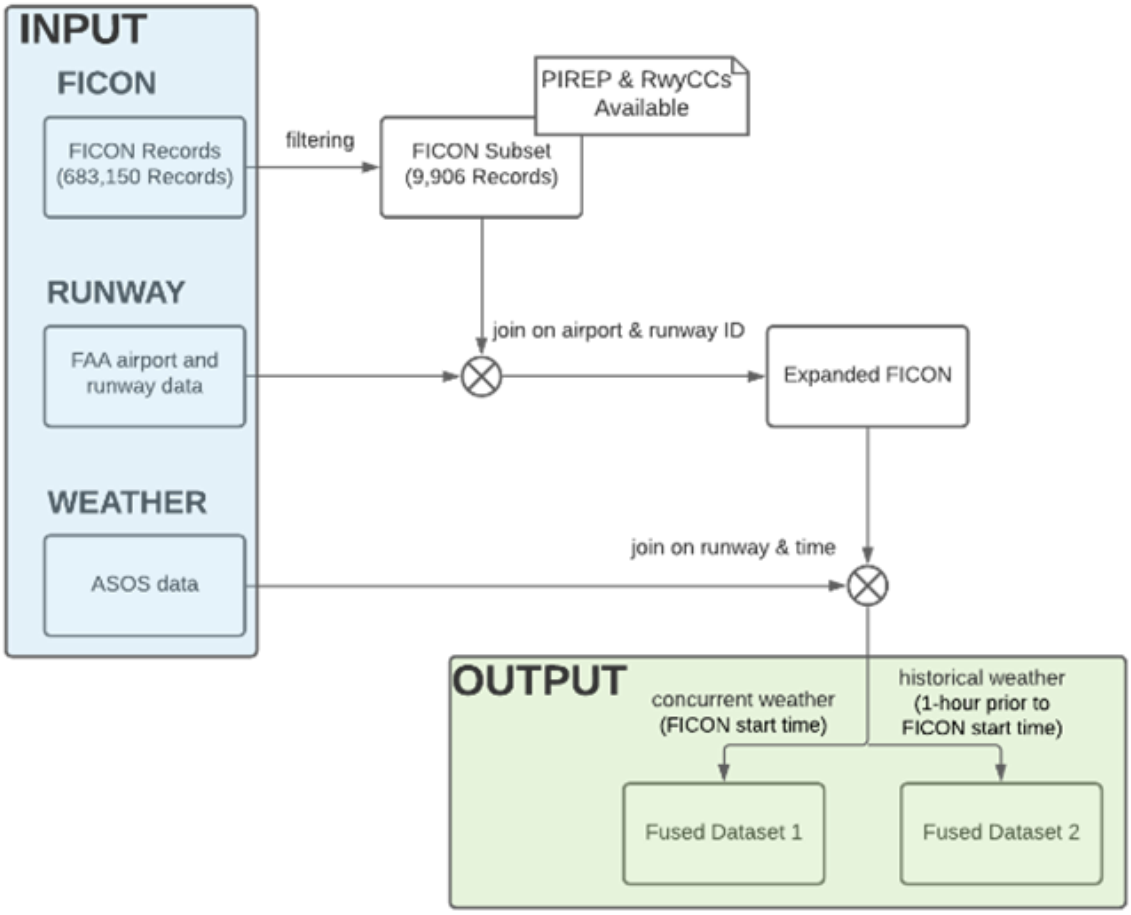


Figure 48. Data fusion for braking and runway condition modeling

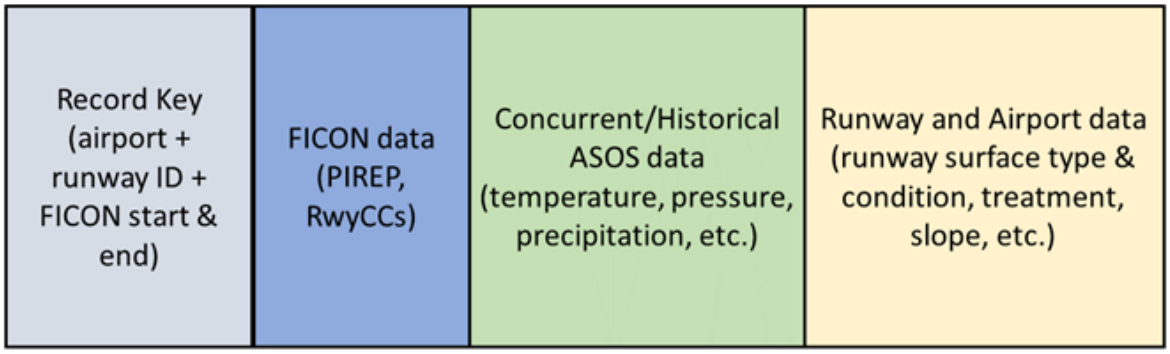


Figure 49. Format of fused dataset #1 and #2

8.1.2 Model overview

Two types of classification models are described in the subsequent sections.

1. PIREP Forecasting (Model #1)

   This model aims to infer PIREPs based on RwyCCs, weather data, and runway characteristics. In practice, it can be used to forecast a PIREP before the aircraft lands and provide the results to pilots to enhance situational awareness prior to landing. Figure 50 provides an overview of this model.

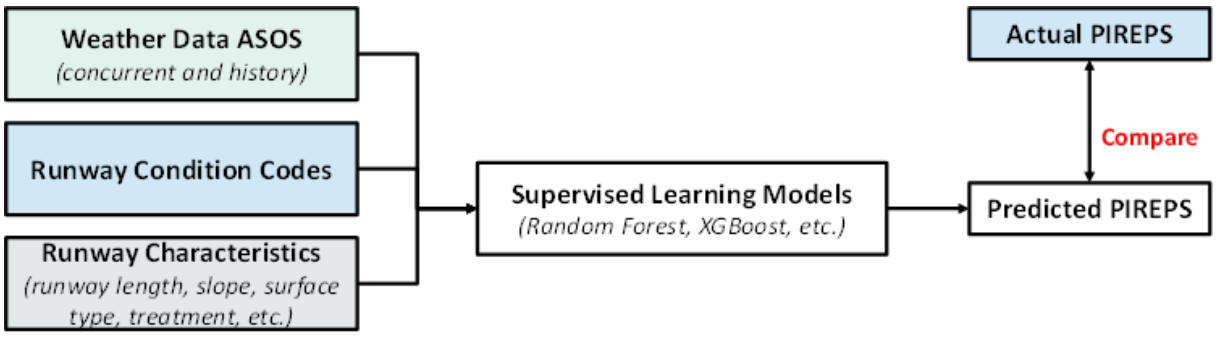


Figure 50. PIREP forecasting model (Model #1)

2. RwyCC Assessment (Model #2)

   This model aims to infer RwyCCs based on weather data, PIREPs, and other runway metadata. In practice, it can be used to assess RwyCCs and compare them against the RwyCCs that are manually assigned to determine how conservative the assignments are compared to pilot reports. Figure 51 provides an overview of this model.

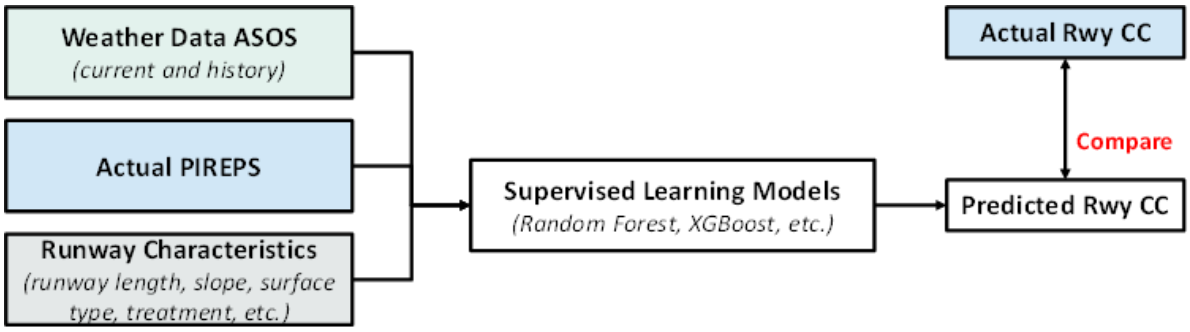


Figure 51. RwyCC assessment model (Model #2)

8.1.3 Implementation

Data preprocessing, including data cleaning, encoding, and feature engineering are conducted first to generate and clean the input data (training and testing sets) required for the machine learning models. Then, random forest and XGBoost are explored to build Models #1 and #2.

- Data cleaning

  Many records in the fused Datasets #1 and #2 are missing either weather or runway characteristics information. This is because the FICON data has a much broader coverage of airports than the ASOS and the FAA runway and airport data. These records which are missing data are not considered further since weather and runway characteristics are indispensable for building machine learning models and it is too difficult to recover these missing data. This step results in 4,499 records (from the original 683,145 FICON records, with and without PIREPs), which have complete information about the PIREPs, RwyCCs, weather, and runway characteristics.

- Data encoding

  The fused datasets contain both numerical and textual values. However, machine learning algorithms typically take only numerical values as inputs. Thus, data encoding is used to convert the textual data into numbers.

  The textual data in the fused datasets is of two types, ordinal and nominal, and two strategies are used to encode these two different types. The ordinal data (such as PIREPs and runway surface condition) has a natural order, so the string values are simply converted into a sequence of integers. The nominal data (such as precipitation type and runway surface type) does not have a natural order, so one-hot encoding is used to convert the nominal data into numbers. Multi-valued one-hot encoding is used for the precipitation type.

  For the precipitation type, there exists several labels that are not explicitly described in the ASOS documents, such as “--" and “?0”. Therefore, these labels are merged into one “unknown” label. Thus, the precipitation type has eight labels: ‘unknown’, ‘NP’ (no precipitation), ‘R-’ (light rain), ‘R’ (rain), ‘R+’ (heavy rain), ‘S-’ (light snow), ‘S’ (snow), ‘S+’ (heavy snow). Data analysis indicates that snow is the most common precipitation type in the dataset. This is because the FICON data used is collected mostly from November to February (during the wintertime). Furthermore, for rain and snow cases, multi-valued one-hot encoding is used in which a sequence of integers represents

the rain and snow precipitation types, and larger values indicate higher intensity of precipitations.

The multi-valued one-hot encoding effectively converts the original precipitation type (represented as a string) into integer numbers. This can mitigate the increase of dimensionality caused by a simple one-hot encoding. Table 9 shows the results of the multi-valued one-hot encoding for the precipitation type (PType).

Table 9. Multi-valued One-hot Encoding Results for Precipitation Type

PType_NP	**PType_Unknown**	**PType_Rain**	**PType_Snow[a]**
0	1	0	0
0	0	2	0
0	0	0	1
0	0	0	3
1	0	0	0
0	0	0	2

a. In this column, "0" means no snow, "1" means "S-", "2" means "S", and "3" means "S+".

- Feature engineering

  An increase in the dimensionality of the input data to a machine learning algorithm would result in an exponential increase in computational efforts. This phenomenon is known as the "curse of dimensionality" (Sheridan, et al., 2020; Verleysen & François, 2005). Consequently, the dimensionality of the problem is reduced by removing redundant information from the input dataset. The multi-valued one-hot encoding applied to the precipitation data is one way of reducing the dimensionality of the problem.

  Data analysis has shown that although each FICON record has three elements for runway condition codes (i.e., RwyCC1, RwyCC2, RwyCC3, representing the three segments/thirds of the runway), ~96% of the records have the same values for these elements. Therefore, only the lowest value among the three is retained for each record. In addition, each record has both a "liftoff elevation" element and a "brake release elevation" element that have very close values.

For convenience, both elements are removed from the dataset, while their mean values are used to create a new feature which represents airport elevation, although the two elevations may be needed in other detailed runway studies.

- Building and selecting machine learning models

  The preprocessed and fused Datasets #1 and #2 are inputs to the machine learning models. Each dataset is divided into two sets such that 80% of the dataset is the training set and the remaining 20% is the testing set. The random forest and the XGBoost machine learning algorithms are then used to build the classification models.

  The scikit-learn library in Python is used to create the eight machine learning models (Pedregosa, et al., 2011). Scikit-learn is a powerful machine learning toolbox that provides frameworks for researchers to efficiently build various machine learning models.

  Models #1 and #2 are two multiclass classification models, where Model #1 classifies PIREPs, and Model #2 classifies RwyCCs. Originally, a PIREP has six textual labels, which are then converted into integers from 1 to 6, where a larger value indicates a better braking action. A RwyCC has six integer labels, ranged from 1 to 6, where a larger value indicates a drier/safer runway condition.

  For each model, the choice of hyper-parameters can significantly affect the performance of the classification algorithm. Therefore, a grid-search algorithm is used to optimize the model performance, where the algorithm exhaustively generates candidates from a grid of parameter values specified by the user and performs experiments on every resulting combination to determine which one leads to the optimal objective value (Belete & Huchaiah, 2021). A 5-fold cross-validation approach is used during the grid-search algorithm to ensure the robustness of the resulting machine learning models.

  The random forest model has four main hyper-parameters (Ali, Khan, Ahmad, & Maqsood, 2012): *n_estimators* describing the number of trees in the classifier; *max_depth* describing the maximum height up to which the trees inside the forest can grow; *min_samples_leaf* describing the minimum number of samples that a leaf node must hold after getting split, which helps to reduce overfitting; and *min_samples_split* describing the minimum number of samples a tree node must hold to split into further nodes.

  The XGBoost model also has four main hyper-parameters (Chen & Guestrin, 2016): *learning rate* which represents the decrease in the step size between iterations to prevent

overfitting; *subsample* which represents the fraction of observations to be randomly sampled for each tree; *max_depth* which represents the maximum depth of individual trees; and *min_child_weight* which represents the minimum sum of weights of all observations required in a child, which can be used as an indirect way to control the depths of trees.

Before the execution of the grid-search algorithm, the range of each hyper-parameter and its search step size need to be defined. The range should be sufficiently large for the search space to contain the optimal result, but not too large to avoid long run times. Thus, we use a heuristic approach to define the range:

- Start from baseline values found in the literature.
- Alter the hyper-parameter values one at a time and observe the resulting change in the objective function.
- Select hyper-parameter ranges of values that correspond to reasonable variations in the objective function.

The output of this process is a complete search space for a particular machine learning model. As mentioned previously, we create two classification models, Model #1 and #2, using fused Datasets #1 and #2 independently for each model. Additionally, two machine learning algorithms are considered to build each model.

This results in a total of eight optimized models, each characterized by a unique combination of classification task, dataset, and machine learning algorithm. Table 10 displays an example search space for Model #1 based on the XGBoost algorithm and the fused Dataset #1.

Table 10. Example of Hyper-parameters Search Range

Hyper-parameter	**Range**	**Step size**
Learning_rate	[0.2, 0.4]	0.05
Max_depth	[10, 20]	2
Min_child_weight	[1, 2, 4, 8]	N/A[a]
Subsample	[0.5, 1]	0.1

a. For this hyper-parameter, the values to be explored are listed in "Range", so there is no step size.

8.1.4 Results

8.1.4.1 Model performance metrics

To evaluate the performance of the machine learning models, we use two reference metrics widely found in the literature: *accuracy* and *F1 score* (Yacouby & Axman, 2020). "Accuracy" measures the number of correctly identified cases and is defined in Equation 2, "F1 score" is defined as the harmonic mean of "precision" and "recall", as defined in Equation 3 and Equation 4. It measures the number of incorrectly classified cases and is provided in Equation 5. Both the "accuracy" and "F1 score" metrics are expected to be large (with a maximum value of 1) for machine learning models with good performance.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative} \quad (2)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

In addition, confusion matrices are used to provide more details on the multi-classification results. Each model also has a feature importance plot showing how well the model represents the input data.

Analyses on the classification results have shown that the type of weather data considered (concurrent in fused Dataset #1 versus historical in Dataset #2) does not significantly affect the model performance. In addition, the optimized random forest and XGBoost models all achieve excellent performance on their training sets (with accuracy and F1 score ≥ 0.98). While this indicates overfitting, experiments have shown that no further regularization can be done. As a result, the rest of the section will be focusing on comparing the results of four classification models applied to their testing sets: Model #1 and Model #2 based on either random forest or XGBoost, with fused Dataset #1 in each case.

8.1.4.2 Discussion of weather data timeliness

The only time information available in the FICON data corresponds to when the FICON records are being issued. Therefore, the weather data, both concurrent and historical, is with respect to the FICON issuance time. The actual landing operation could have happened any time during which the FICON was active.

This works well for Model #2, since it is used to infer RwyCCs, which are typically assigned at the FICON issuance time. However, the weather data for Model #1 should ideally be with respect to the aircraft landing operation. Unfortunately, the aircraft landing operation time is unavailable in the current data, so the weather data associated with the FICON issuance time is used for both models.

8.1.4.3 Model #1

Figure 52(a) and Figure 52(b) demonstrate the optimized performance of Model #1 based on the random forest and XGBoost machine learning algorithms respectively. The optimized Model #1 with random forest has the following hyper-parameter set:

- number of trees = 80
- max_depth = 25
- min_samples_leaf = 1
- min_samples_split = 2

The optimized Model #1 with XGBoost has the following hyper-parameter set:

- learning rate = 0.3
- max_depth = 18
- min_child_weight = 0.5
- subsample = 0.8

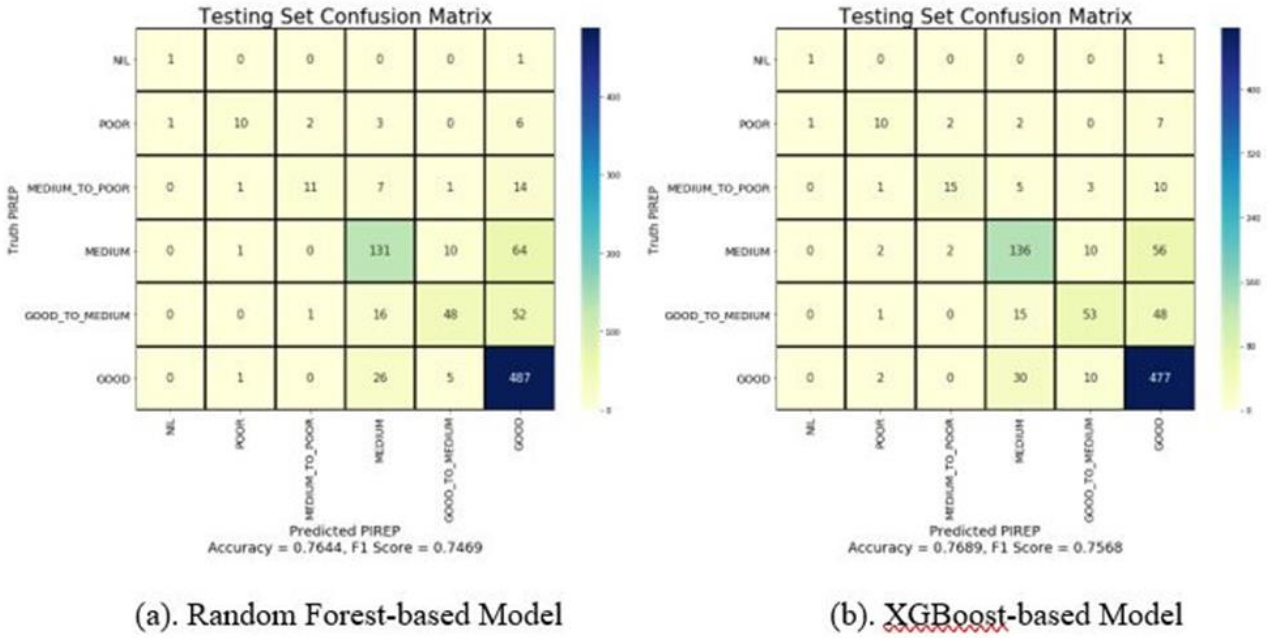
We observe that the two models have very similar performance in terms of accuracy and F1 score.

In a confusion matrix, each cell represents the number of samples that share the same classified value and the same truth value. For instance, in Figure 52(a), the cell on the second row and the last column with a value of 6 indicates that there are 6 samples that are classified to be “GOOD” by the model, when in fact they are “POOR”. Additionally, the six cells on the matrix diagonal represent samples that are correctly classified by the model. In turn, the cells not on the matrix

diagonal represent samples that are misclassified. Moreover, the farther away a cell is from the diagonal, the worse the classification is in that cell.

Finally, the matrix "upper triangle" (excluding the diagonal) represents the "over-estimated" cases, for which the forecasted PIREPs are better than the actual PIREPs, while the matrix "lower triangle" represents the "under-estimated" cases, for which the forecasted PIREPs are more conservative than the actual PIREPs. In practice, under-estimation is preferred to over-estimation, since an overly optimistic signal may mislead pilots and lead to potentially dangerous conditions.

By comparing the individual cells in the two matrices depicted in Figure 52, we observe that the two models achieve the same level of performance in terms of confusion matrix.



(a). Random Forest-based Model (b). XGBoost-based Model

Figure 52. Classification results for Model #1

Figure 53(a) and Figure 53(b) depict the feature importance for Model #1 based on random forest and on XGBoost respectively. Feature importance describes which variables (in our dataset, which elements) are most relevant to evaluate the objective function. The higher the importance of a feature, the more the variable plays a role in the value of the objective function when evaluated from the machine learning model considered.

We observe that the random forest-based model ranks "air pressure" as the most important feature, followed by "wind direction" and "dew/air temperature", while "RwyCC" is ranked at

the 5th position. In addition, the variance of feature importance scores is large, and features related to precipitation, runway surface, and runway treatment have very small importance scores, which indicates the model interprets them as having little correlations with PIREPs. However, the XGBoost-based model ranks "RwyCC" as the most important feature to evaluate PIREPs and produces a more balanced distribution for other feature importance scores.

It is reasonable for "RwyCC" to be ranked as the most important feature with a significantly large score, since it is a holistic representation of runway condition. It would also make more sense to assign the other features relatively large scores, especially those related to runway surface and treatment. Overall, the feature importance of the XGBoost-based model is more reasonable in the context of evaluating PIREPs. Therefore, although the random forest-based model achieves the same level of performance in terms of accuracy and F1 score, XGBoost is considered as the most appropriate machine learning algorithm to build Model #1.

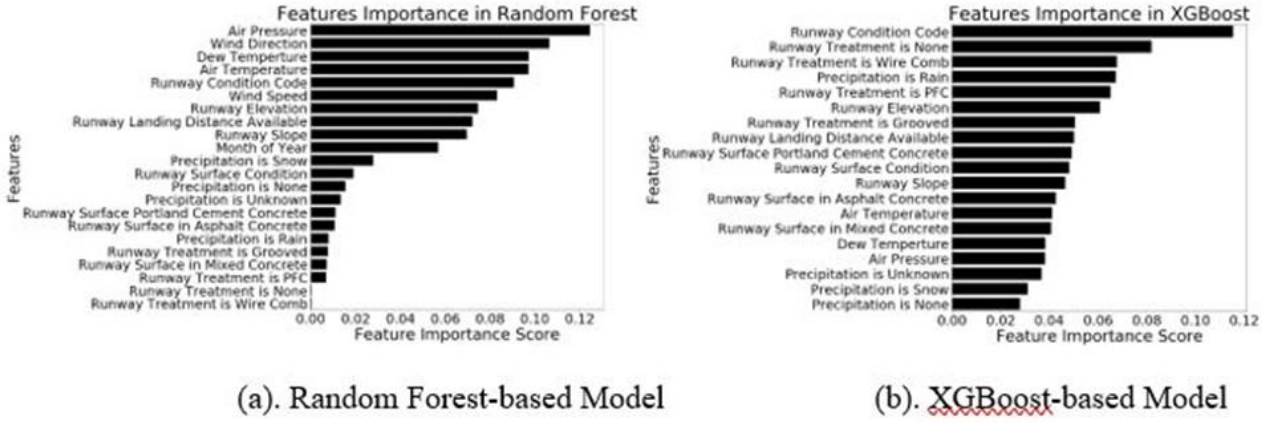


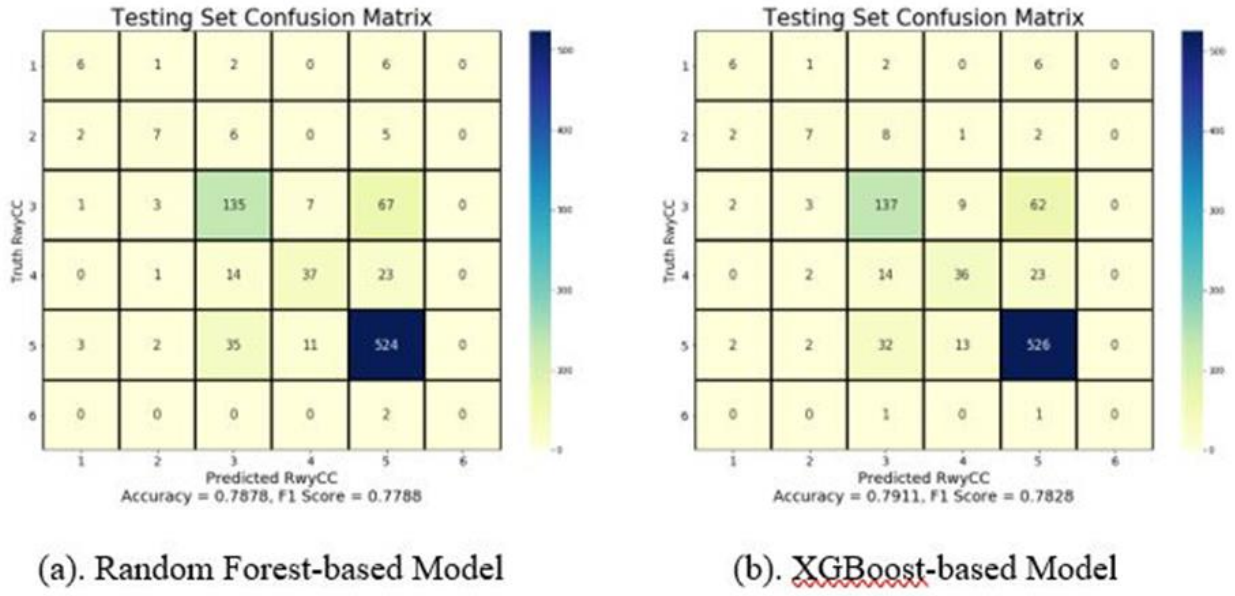
Figure 53. Feature importance in Model #1

8.1.4.4 Model #2

Figure 54(a) and Figure 54(b) demonstrate the performance of Model #2 based on random forest and XGBoost respectively. We observe that the two models achieve the same level of performance in terms of accuracy, F1 score, and confusion matrix.

Overall, the performance of Model #2 is slightly better compared to Model #1. The accuracy and F1 score are around 0.78 for Model #2, while they are around 0.75 for Model #1. This may be due to the timing of the weather data, since it is concurrent with the FICON issuance time (when RwyCCs are assigned), and therefore it is closely related to the issuance of RwyCCs. However, a PIREP is reported by the pilot after the aircraft has landed, and thus the weather data should ideally be concurrent with the aircraft landing time when evaluating PIREPs. Unfortunately, the

aircraft landing time is not available, so we use weather data concurrent with the FICON issuance time for both Model #1 and Model #2. This can result in a more accurate classification for RwyCCs than for PIREPs.



(a). Random Forest-based Model (b). XGBoost-based Model

Figure 54. Classification results for Model #2

The feature importance analysis is also carried out for the aforementioned two models. We observe that the XGBoost-based model again results in a better interpretation of the input data, since it ranks PIREPs as the most important feature, followed by "runway treatment type", "surface type and condition", and "air temperature", which follows the common sense of physics.

The results for Model #2 are similar to those for Model #1 in terms of model performance and data interpretation. For Model #1, the PIREP is the objective function and the RwyCC is a feature, whereas for Model #2, the RwyCC becomes the objective function and the PIREP a feature. The other features remain the same. Therefore, the input datasets for the two models are very similar to each other.

8.1.4.5 Summary of results

The results are summarized as follows:

- Tree-based machine learning algorithms, namely random forest and XGBoost, are suitable for building Models #1 and #2.

- Although random forest achieves the same level of performance as XGBoost, the interpretation of the input data by the XGBoost-based model is more reasonable, which makes XGBoost the most suitable algorithm for both Model #1 and Model #2.
- The performance of Model #2 is consistently better than that of Model #1 due to the timeliness of weather data.
- The results do not change significantly when switching between concurrent and historical weather data.

8.1.5 Conclusions

Supervised machine learning models are built by using an integrated dataset resulting from the fusion of various relevant data sources required to infer PIREPs and RwyCCs. The models, especially those built using the XGBoost algorithm, have reasonably good performance, and thus demonstrate that it is possible to build classification models for aircraft braking performance and runway condition predictions. The classification model for forecasting PIREPs can be used to enhance the situational awareness of pilots prior to actual landing operations, whereas the model for predicting RwyCCs can be used to compare machine learning results with manual assignments to study the conservativism in a retrospective manner.

8.2 Aircraft acceleration evaluation

The goal of this task is to evaluate aircraft braking performance based on flight data. The aircraft braking performance is represented by aircraft longitudinal acceleration. Both the longitudinal acceleration and the relevant flight data come from FOQA data described in Section 3.1.1.

8.2.1 Overview

The evaluation of aircraft longitudinal acceleration can be considered a regression task. Supervised machine learning models such as Random Forest, XGBoost, and Artificial Neural Networks (ANN) (Abadal, Jain, Guirado, López-Alonso, & Alarcón, 2021), can be utilized to solve the regression problem. Random Forest and XGBoost are used for classification purposes in Section 8.1 and for regression in this task.

The target for the regression is aircraft longitudinal acceleration, and the following nine columns in the FOQA data are selected as the features for regression:

- N1 of the left engine (%)
- Speed brake position (degree)

- Airspeed (knots)
- Ground speed (knots)
- Brake pressure (psi)
- Gross weight (lb)
- Total fuel quantity (lb)
- Ground track distance to touchdown (nm)
- Ground track distance to threshold (nm)

8.2.2 Data description

The data used in this task is a subset of the available FOQA dataset. In total, 217 landing flights of the NB-B2/3 airframe at the Harry Reid International Airport (LAS) in Las Vegas are collected. All flights occurred on dry runways. These flights are all processed to include only the segment where spoilers and thrust reversers are activated, which results in 10-20 seconds for each flight. Consequently, the aircraft performance considered in this task is the maximum braking during landing operations. The final dataset contains 2,440 records, each corresponding to a particular timestamp in a landing event.

Although the size of the dataset seems to be small, considering that the scenario being studied is quite specific (i.e., NB-B2/3 airframe, LAS airport, and dry runways only), it would still be sufficient for the supervised machine learning tasks. In addition, although the statistical model built in this task is based on dry runways, it can be generalized to contaminated runways. This can be done by following the same modeling framework and replacing the input data with flights on contaminated runways.

8.2.3 Implementation

- Random Forest and XGBoost Models

  The scikit-learn library in Python is used to implement the Random Forest and XGBoost models. A grid-search with 5-fold cross-validation is used to tune the hyperparameters and obtain the optimized models. The search space of both models is summarized in Table 11.

Table 11. Hyperparameter search space for random forest and XGBoost

Random Forest		**XGBoost**	
Hyperparameter	**Search Space**	**Hyperparameter**	**Search Space**
Estimators	[5, 20, 40, 60, 80, 100, 120, 140]	Learning rate	[0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4]
Max tree depth	[5, 10, 15, 20, 25, 30]	Max tree depth	[4, 6, 8, 10, 12, 14, 16, 18, 20]
Min samples for splitting	[2, 4, 6, 8]	Min child weight	[0.5, 1, 2, 4]
Min samples in leaf	[1, 2, 4, 8]	Subsample portion	[0.5, 0.6, 0.7, 0.8, 0.9, 1]

- Artificial Neural Network (ANN)

  A multi-layer perceptron (MLP) is chosen to be the architecture of the ANN model for regression. The Tensorflow-Keras framework is used for implementing the model.

  The hyperparameters of the ANN model are as follows:

  - Number of layers
  - Number of neurons in each layer
  - Learning rate of the optimizer
  - Number of epochs
  - Dropout rate of each layer

  In this case, an *epoch* is an iteration during the training process, and dropout is a technique used in ANN to combat overfitting in which some neurons in the neural network are chosen to be ignored with a given probability during the training process.

  Training an ANN model typically takes significantly more time than training a Random Forest or an XGBoost model. Therefore, GridSearch is not used for hyperparameter tuning in the case of the ANN model. Instead, the Hyperband searching technique (Li, Jamieson, DeSalvo, Rostamizadeh, & Talwalkar, 2017) is used where the algorithm attempts multiple configurations of hyperparameters for a few epochs and proceeds only with those configurations that have the best performance. Hyperband conducts “adaptive” search and thus results in a much shorter search time while maintaining a good result

compared to the exhaustive search. The hyperparameter search space for the ANN model are summarized in Table 12.

Table 12. Hyperparameter search space for ANN

Number of Layers	**Number of Neurons in each Layer**	**Number of Epochs**	**Dropout Rate**	**Learning Rate**
1	[8:160, 8]	1-150	[0:0.8, 0.1]	[1e-5, 1e-4, 1e-3, 1e-2]
2	[16:160, 16] [8:80, 8]	1-150	[0:0.8, 0.1]	[1e-5, 1e-4, 1e-3, 1e-2]
3	[16:160, 16] [8:80, 8] [4:40, 4]	1-150	[0:0.8, 0.1]	[1e-5, 1e-4, 1e-3, 1e-2]
4	[16:160, 16] [16:80, 16] [8:40, 8] [4:20, 4]	1-150	[0:0.8, 0.1]	[1e-5, 1e-4, 1e-3, 1e-2]

8.2.4 Results

8.2.4.1 Model performance metrics

To assess the performance of the regression models, the R^2 value is used, which quantitively measures how close the data is to the fitted regression line as shown in Equation 6. Additionally, predicted vs. actual plot and residual plot are used to provide more details of regression results. A good fit is expected to have the residual following a normal distribution around 0.

$$R^2 = 1 - \frac{Residual\ sum\ of\ squares}{Total\ sum\ of\ squares} \quad 6$$

8.2.4.2 Model comparison

The optimized Random Forest model has the following hyperparameters:

- estimators = 120
- max tree depth = 15
- min samples in leaf = 1

- min samples for splitting = 6

The optimized XGBoost model has the following hyperparameters:

- learning rate = 0.15
- max tree depth = 6
- min child weight = 4
- subsample portion in each round = 0.9

The optimized ANN has the following hyperparameters:

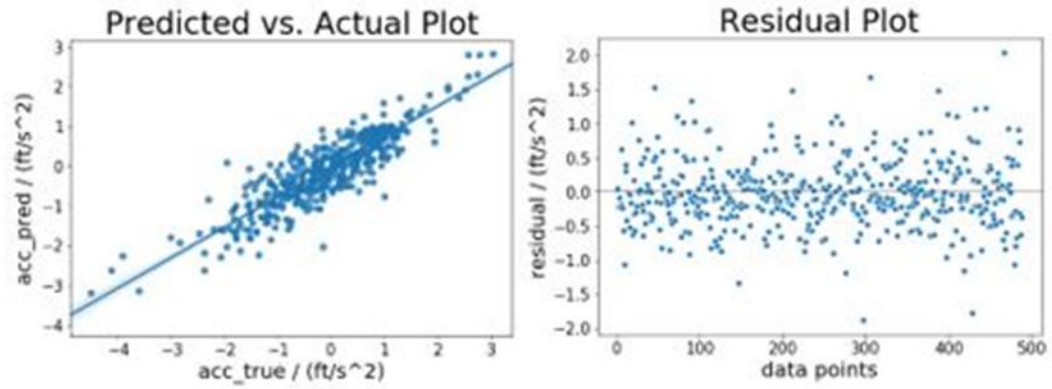
- number of layers = 2
- number of neurons at each layer = 144, 64
- dropout rate at each layer = 0.1, 0
- number of epochs = 539
- learning rate = 10^{-4}

The R^2 values on the testing set of the three optimized models are as follows:

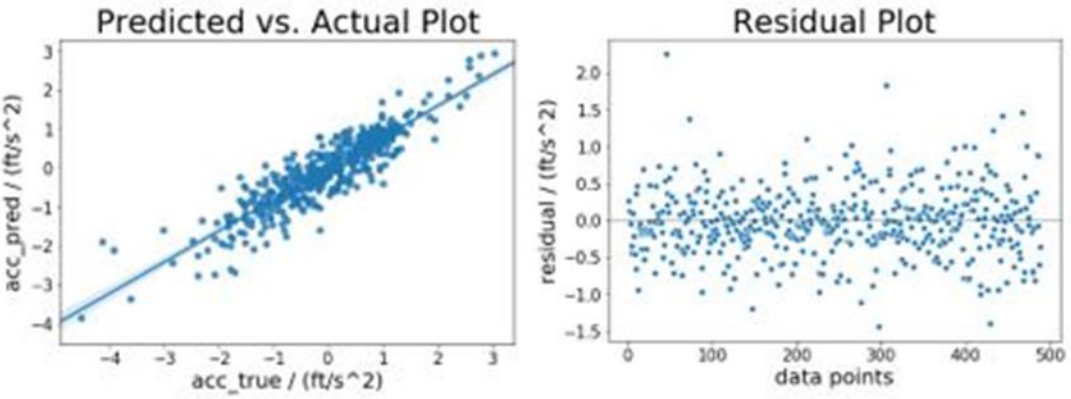
- Random Forest $R^2 = 0.70$
- XGBoost $R^2 = 0.76$
- ANN $R^2 = 0.80$

It is worth mentioning that for the ANN model, multiple hyperparameter settings can achieve the same result. For instance, another ANN model was observed to have R2 = 0.80 with 3 layers. We have selected to use the one shown above.

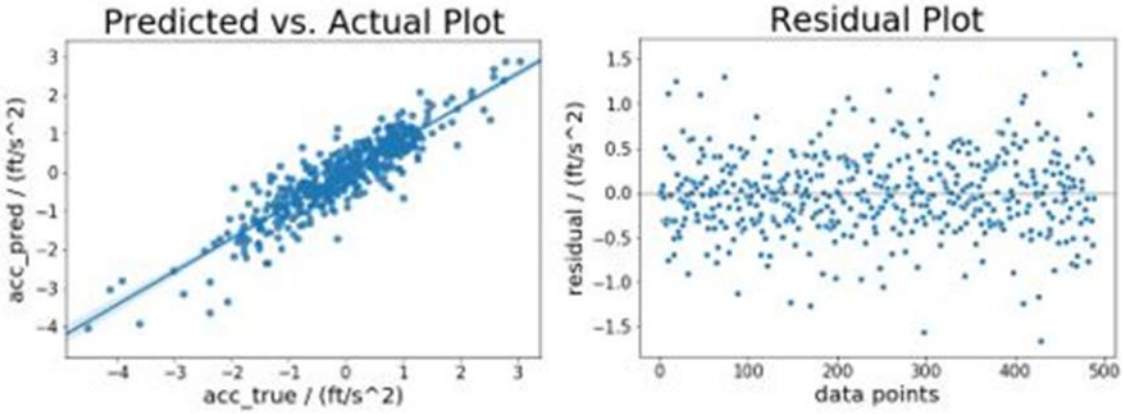
Results show that the ANN model outperforms Random Forest and XGBoost in the regression task in terms of the R^2 value. It can be observed in Figure 55 that the residuals in all three models have a roughly normal distribution around the x-axis, which indicates that all three models have a reasonably unbiased fit.



(a). Random Forest Results



(b). XGBoost Results



(c). ANN Results

Figure 55. Regression results for random forest, XGBoost, and ANN models

Based on the aforementioned observations, we can conclude that the ANN model has the best overall performance.

8.2.4.3 Relationship between features and target

Study regarding the relationship between several features and the target is conducted to explore the interpretation of the ANN model. The study is carried out in the following manner:

- Study the features one at a time, while keeping the other eight fixed by taking the mean values, which results in 9 experiments in total.
- In each experiment, take as the input the feature values from the testing set and get the target values through the ANN model.

This study has the ability to demonstrate the ANN model interpretation of the correlations between the features and the target. It can also show the importance of each feature in the model, since more importance features would lead to larger variations in the target.

Figure 56 provides the definition of ground track distance in the FOQA dataset. According to the definition, ground track distance to touchdown and threshold will both have negative values in the segment we are interested in (spoiler deployed and thrust reversers activated), while their absolute values increase as the landing proceeds.

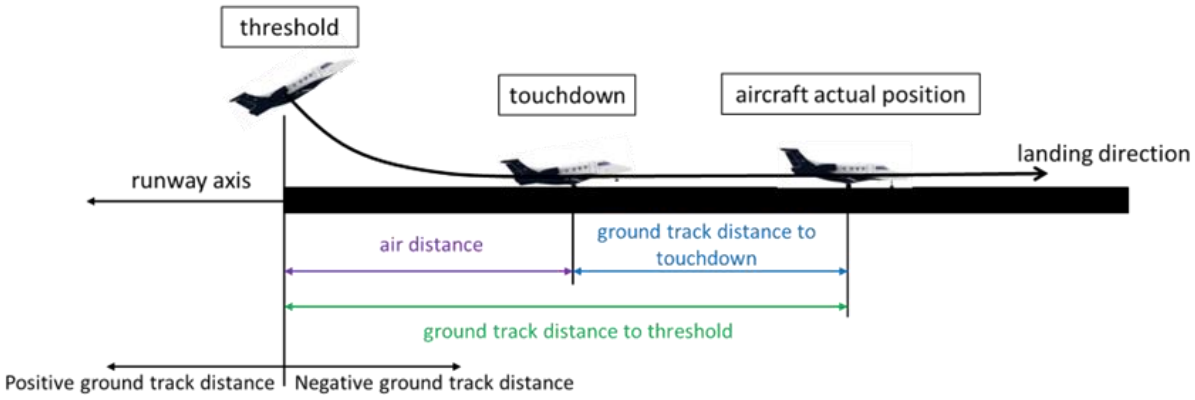


Figure 56. Ground track distance definition in the FOQA dataset

Figure 57(a) to Figure 57(i) are scatter plots showing the relationships between the features and the target. Table 13 quantitatively summarizes how each feature induces the observed variations in the target, which indicates feature importance.

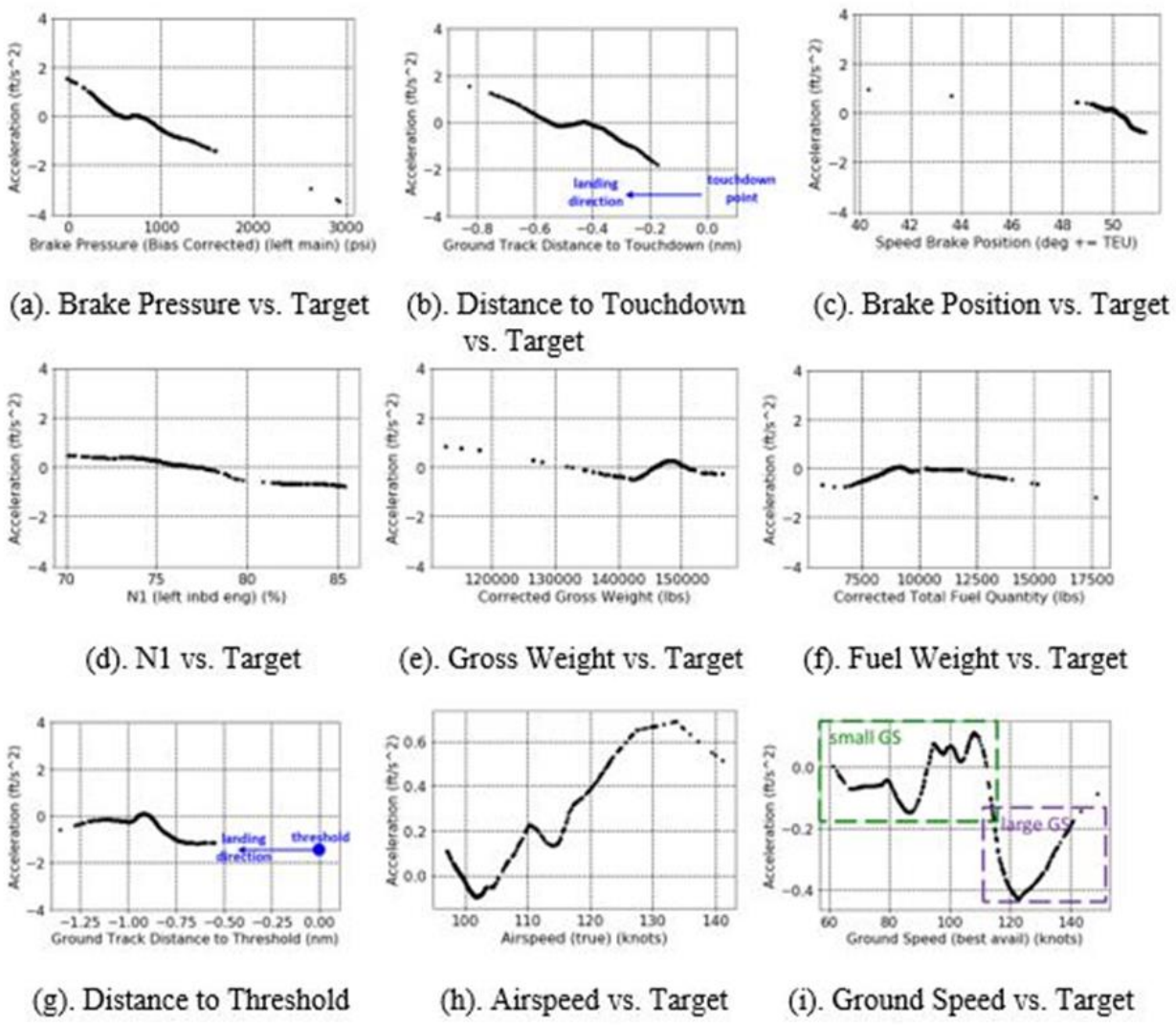


Figure 57. Features vs. target scatter plots

Table 13. Feature range vs. induced target range

Feature Name	**Feature Unit**	**Feature Range**	**Induced Variance in Acceleration (ft/s^2)**	**Importance Rank**	**Correlation to Braking**
Brake Pressure	Psi	[-16.7, 2924.8]	5.00	1	+
Ground Track Distance to Touchdown	Nm	[-0.17, -0.83]	3.36	2	+

Feature Name	Feature Unit	Feature Range	Induced Variance in Acceleration (ft/s^2)	Importance Rank	Correlation to Braking
Speed Brake Position	Degree	[40.4, 51.3]	1.73	3	+
Gross Weight	Lbs	[112,784 -> 156,461]	1.32	4	-/+/-
Ground Track Distance to Threshold	Nm	[-0.55, -1.36]	1.26	5	+
N1 (Left Engine)	%	[70.1, 85.4]	1.25	6	+
Fuel Quantity	Lbs	[5774.7, 17694.0]	1.24	7	+/NON/-
Airspeed	Knots	[97.1, 141.2]	0.78	8	-
Ground Speed	Knots	[61.3, 148.7]	0.54	9	+

It can be observed in Table 13 that brake pressure, speed brake position, and ground distance to touchdown are the three most important features when evaluating acceleration, since any variation in these parameters cause the largest variation in acceleration. In the meantime, airspeed and ground speed are the least important features. These are expected behaviors because brake pressure and brake position are directly related to the braking action.

It is worth mentioning that in Figure 57, a smaller value in acceleration indicates a greater braking, since a braking action causes a decrease in acceleration. It can be observed from Figure 57(a), (c), and (d) that brake pressure, speed brake position, and engine N1 % have positive correlation to braking, which reflects real operations correctly. In Figure 57(b), braking is observed to be the greatest after the aircraft touchdown point and then gradually diminishes. A similar behavior can be observed in Figure 57(g). These are also reasonable behaviors since they reflect the situation where pilots would try the hardest to stop the airplane right after the touchdown point and then gradually release the brakes. The braking behavior is more complicated with respect to gross weight and fuel weight, as shown in Figure 57(e) and (f). Finally, it can be seen from Figure 57(h) and (i) that braking is greater when ground speed is larger, while the trend is opposite with respect to airspeed.

8.2.5 Conclusions

In this task, three machine learning models are implemented to evaluate the aircraft braking performance based on relevant flight data during the landing phase. The ANN model turns out to be the best since it achieves the highest R2 value and has an unbiased residual plot. Thus, we conclude that the ANN model is preferable in the regression task. Feature studies are carried out based on the ANN model, and it is shown that the model interpretation of the input features reflects the real operations to a certain degree.

While 0.80 is a reasonable R2 value, the ANN model could still achieve an even better performance. This may be obtained by increasing the size of the input data. In turn, the feature study based on a model with improved performance is also expected to better reflect the real landing operations.

9 Phase II - Degraded Braking Research Roadmap

9.1 Roadmap goals

The goals of the roadmap are twofold:

1. Quantify the conditions and performance factors associated with actual landing distances. This develops a foundation for the analysis of operations on reduced friction/contaminated runways in order to satisfy the FAA vision to go from a qualitative to a data-centric/quantitative approach.
2. Evaluate the following NTSB recommendations (NTSB, 2016):
    - **A-16-23**: Continue to work with industry to develop the technology to outfit transport-category airplanes with equipment and procedures to routinely calculate, record, and convey the airplane braking ability required and/or available to slow or stop the airplane during the landing roll.
        - This specific issue has been partially addressed via the creation of ASTM standard E3266-20 “Standard Guide for Friction-Limited Aircraft Braking Measurements and Reporting” (ASTM, 2020). This standard in conjunction with draft Advisory Circular AC 91-79B covers many of the issues associated with low friction runway conditions - AC 91-79B is a significantly expanded version of AC-91-79A (FAA, 2018). Therefore, A-16-023 will not be the focus of this roadmap.

- **A-16-24:** If the systems described in Safety Recommendation A-16-23 are shown to be technically and operationally feasible, work with operators and the system manufacturers to develop procedures that ensure that airplane-based braking ability results can be readily conveyed to, and easily interpreted by, arriving flight crews, airport operators, air traffic control personnel, and others with a safety need for this information.

    - The reporting of airplane-based braking abilities is discussed in draft AC 91-79B. However, this reporting is limited to current technologies and FAA information distribution systems (i.e., ATIS, NOTAMs, and FICONS). However, comprehensive reporting and information distribution systems for automated reporting from aircraft has yet to be developed. This will be a discussion item for the roadmap.

9.2 Roadmap purpose

The purpose of the roadmap is to depict a gap analysis to meet the aforementioned goals, i.e., to identify *critical paths* for:

- *Research*: data analysis, machine learning techniques investigation and development
- *Data gathering*: amount, accuracy, resolution, source (airport, aircraft, pilot, weather, etc.), synchronization
- *Instrumentation*: sensors required, locations of deployment, conditions of deployment (runway, weather, etc.), platform for deployment
- *Testing, Validation*, etc.

9.3 Roadmap focus areas

The roadmap features four areas of investigation as described below and in Figure 58, and identifies interactions between focus areas and related challenges.

- **Airport:** *Main focus*

  The identified gap is the centralized reporting of the wheel braking data received from the aircraft or the reduction in runway friction, as a function of runway condition, aircraft performance and state at landing, potential weather conditions (temperature, pressure, humidity, ceilings, etc.), and piloting.

- **Aircraft**

The reporting of aircraft wheel braking using an on-aircraft system is currently defined using the ASTM standard E3266-20. Therefore, only critical interdependencies or relationships with other focus areas (airport in particular) will be highlighted on the roadmap.

- **Pilot**

  The research highlighted in the roadmap will address potential gaps in training or standard operating procedures.

- **Weather**

  Research in this area is probably more advanced than in other areas, so only critical interdependencies or relationships with other focus areas will be highlighted on the roadmap, along with potential requirements for data with other accuracy, resolution, synchronization, source, etc.

  Note: weather and FICON data fusion has been developed earlier by the Georgia Tech team as described in other sections of this report.

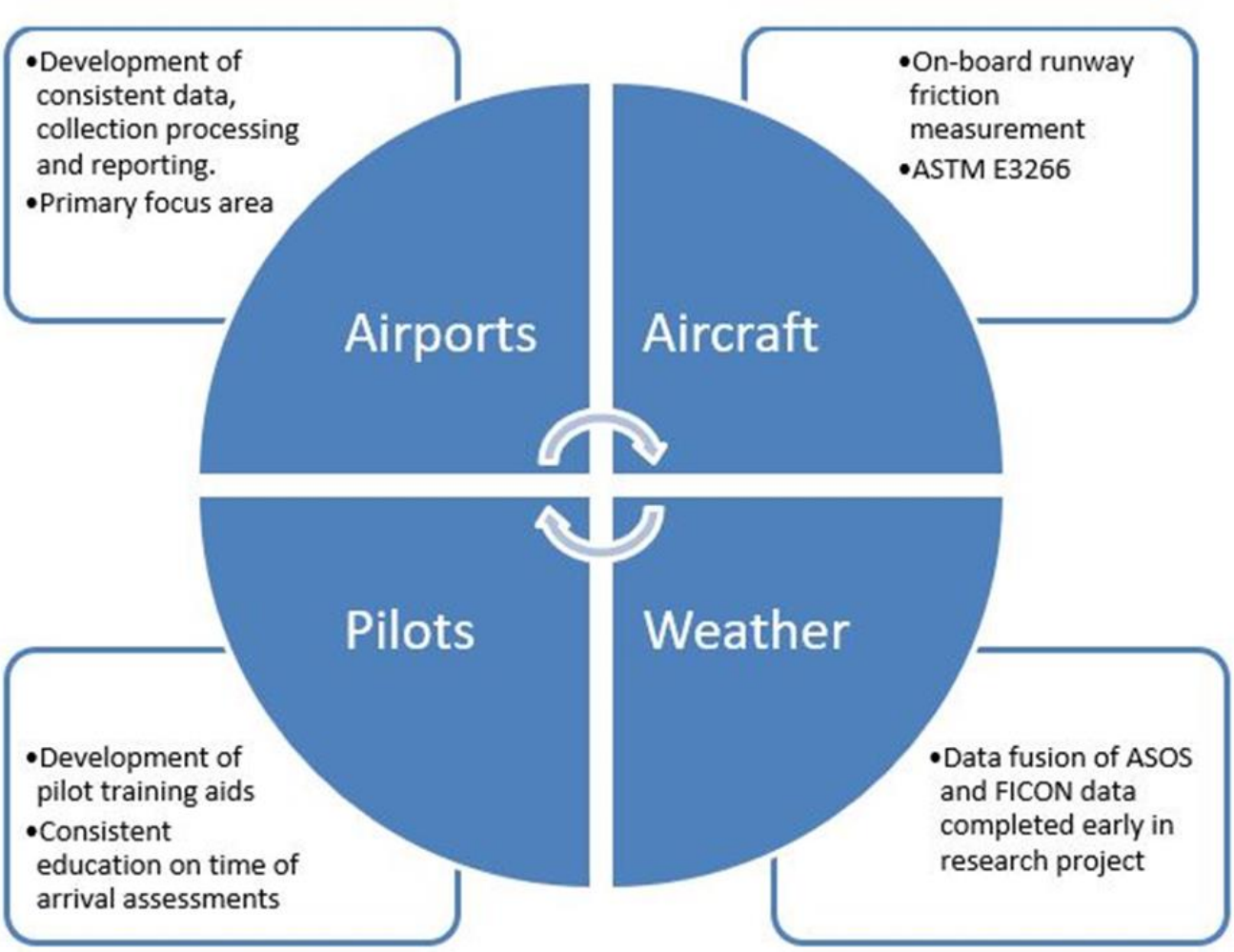


Figure 58. Roadmap focus areas

9.4 Roadmap development process

The roadmap is developed as follows:

- Start with four aforementioned areas, with a focus on **airport**.
- Identify data required to inform goals from various areas of interest.
- Determine if data is available, sources it is available from, and if it is available in appropriate format, at appropriate accuracy, resolution, synchronization with other data sources, etc.
- Identify what sensors/instrumentation would be needed to gather data if not available (e.g., in-runway temperature, depth sensor, etc.).

- Identify technology required to gather data if sensors/instrumentation are not available.
- Determine data gathering process if data not available.
- Suggest high-level data analysis and potential machine learning techniques (clustering, classification, etc.).
- Explore interactions between airport focus area and others.

9.5 Roadmap considerations

In the following section that outlines the roadmap elements, each of the Focus Areas and Focus Items will be subject to some, if not all, of the considerations shown in Figure 59 and their possible interdependencies.

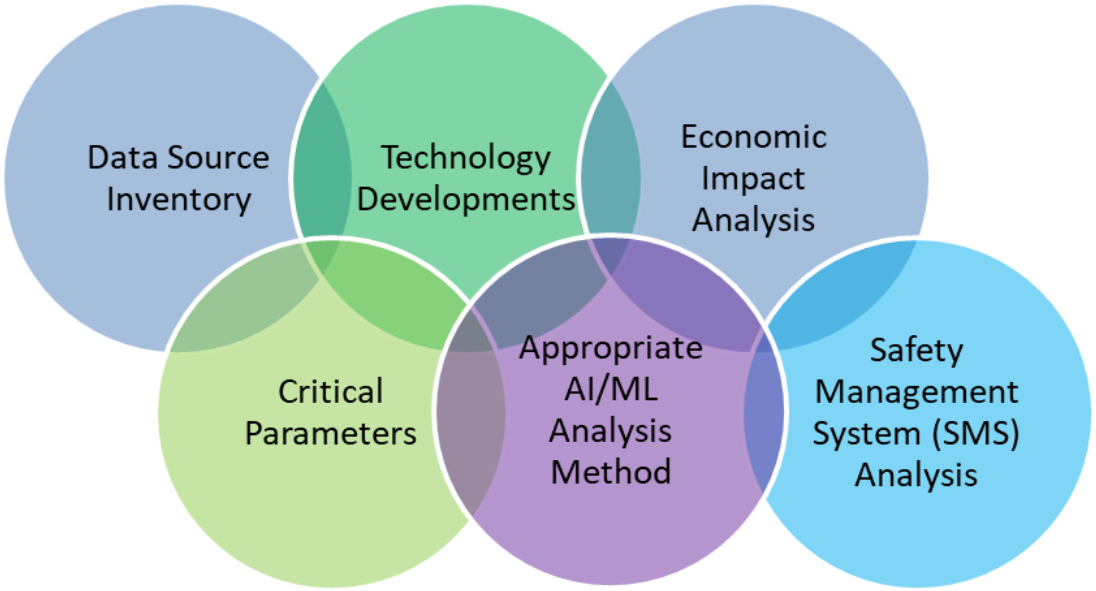


Figure 59. Roadmap considerations

9.6 Roadmap elements

The previous Section 9 paragraphs have provided the macro components and considerations for a runway safety and landing performance research roadmap. This section contains a more granular representation of possible research areas. This material is intended for regulators and researchers that endeavor to create specific research projects. The topics are notional and should be considered a starting point for any potential future research project.

Focus Areas have been previously discussed in Section 9.3 and provide a basic partitioning of the research areas. *Items* were selected based on an iterative review process between the FAA and Georgia Tech Subject matter experts (SMEs) and an exhaustive literature review was not required in this case. An understanding of current best operational practices for airport and airlines provided a context for what may be the most beneficial safety enhancements. Best practices are found in FAA regulatory and operational guidance materials such as AC 91-79A.

Obtaining the *Data Required* is the most challenging component for any specific *Items*. This requires either the support of airline or airport operators to supply data that is collected during normal operations, or a direct testing and data gathering exercise at physical locations. Operator data access is often constrained by Intellectual Property (IP) and commercial competitive concerns. Field tests and data gathering is typically constrained by budgets, manpower resources, and facility access approvals. These types of data gaps and challenges to the process are identified by orange coloring. Recommendation of solutions to those constraints is beyond the scope of the Roadmap proposal and is therefore not addressed in this report.

Lastly, *Bridging the Data Gap* is an analysis and policy proposal exercise. The efforts involved could be significant and lengthy. However, this roadmap being a proposal for future avenues of research at the FAA, many of the issues related to data gathering do not extend to the *Bridging the Gap step*. The administrative and policy changes that may be needed to support future operational practices is not a technical constraint. How these gaps might be addressed is the purview of the applicable regulatory agency such as the FAA.

Figure 60 shows the roadmap legend for the focus areas shown in the subsequent sections as part of the runway safety and landing performance research roadmap.

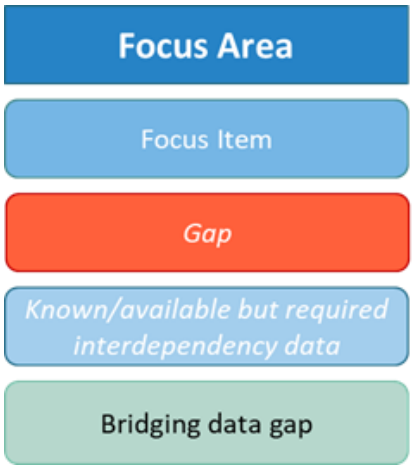


Figure 60. Roadmap legend

9.6.1 Airport focus area

Figure 61 shows the research roadmap for the airport focus area, with the following acronym descriptions.

- RwyCC: Runway Condition Code
- ABAR: Aircraft Braking Action Report
- PIREPs: Pilot Reports
- FICON: Field Condition Report

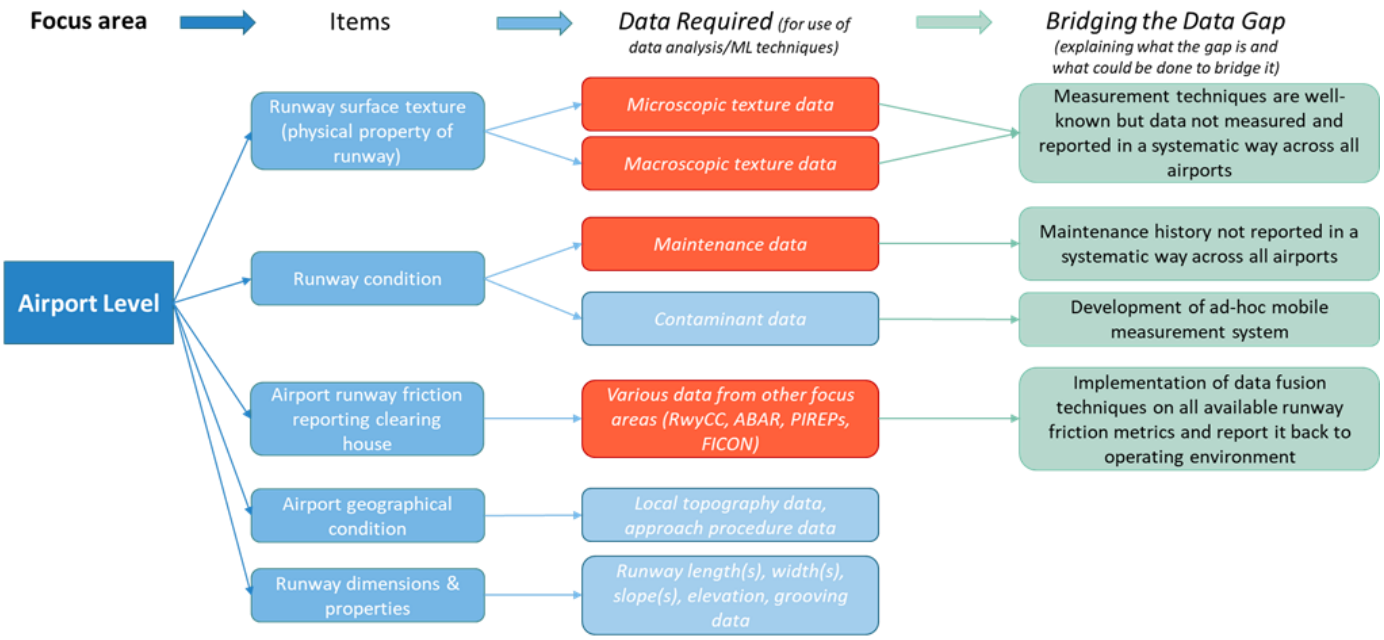


Figure 61. Airport focus area roadmap

9.6.2 Aircraft focus Area

Figure 62 shows the research roadmap for the aircraft focus area, with the following acronym descriptions.

- QAR: Quick Access Recorder
- FOQA: Flight Operations Quality Assurance
- ASIAS: Aviation Safety Information Analysis and Sharing

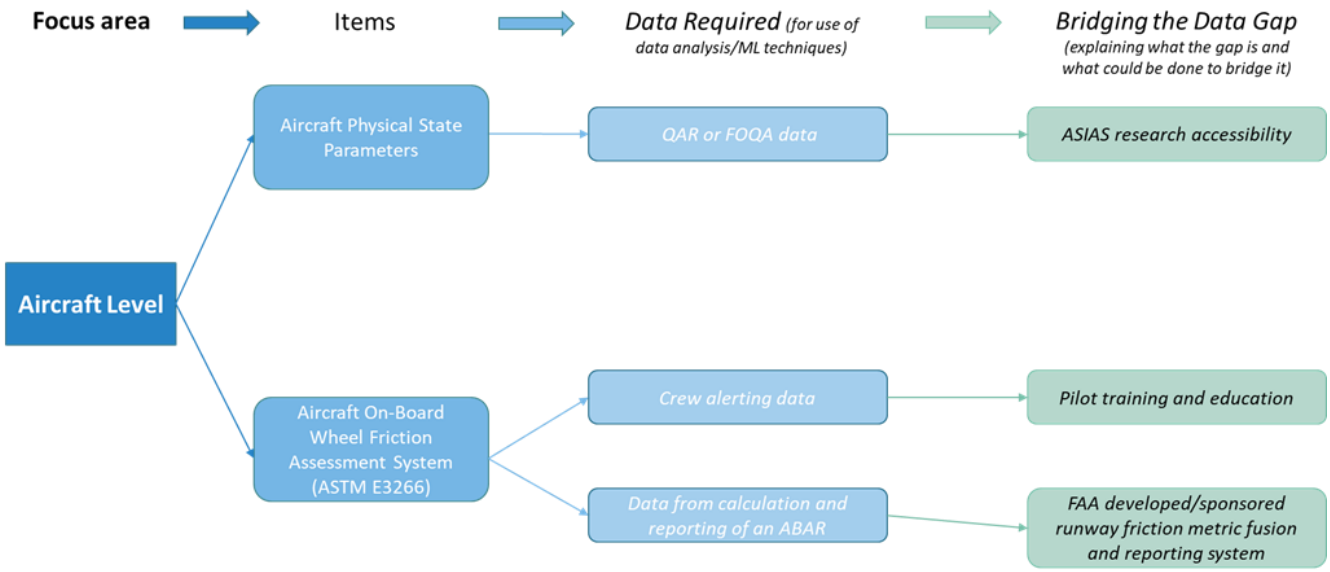


Figure 62. Aircraft focus area roadmap

9.6.3 Pilot focus area

Figure 63 shows the research roadmap for the pilot focus area, with the following acronym descriptions.

- PIREPs: Pilot Reports
- ABAR: Aircraft Braking Action Report
- ATC: Air Traffic Control

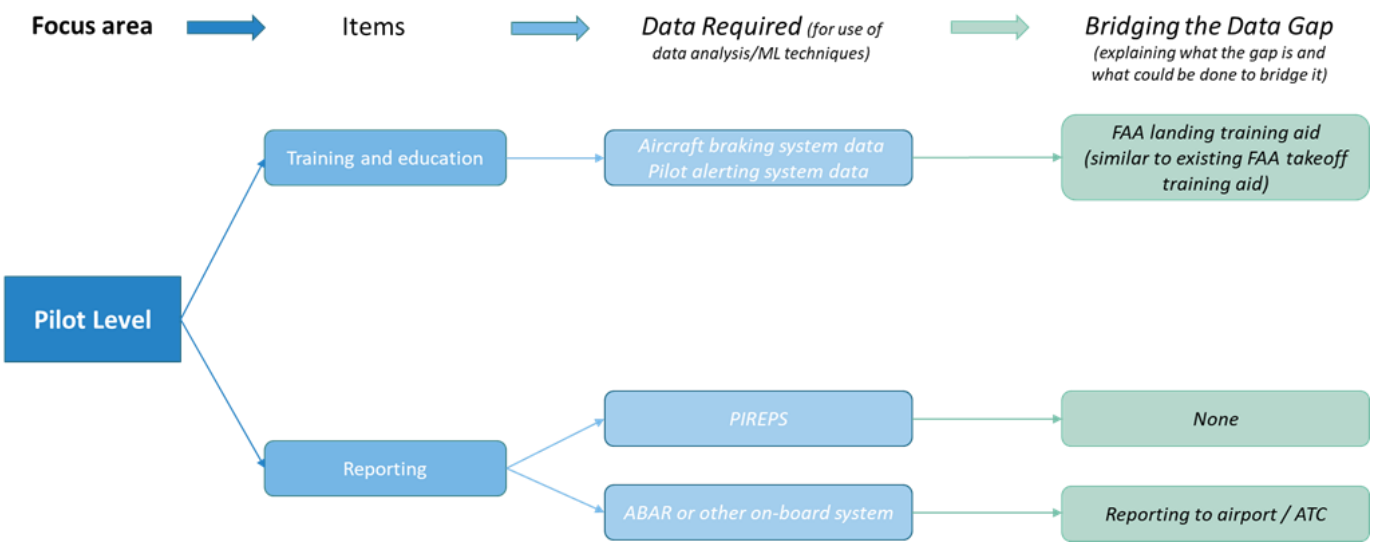


Figure 63. Pilot focus area roadmap

9.6.4 Weather focus area

Figure 64 shows the research roadmap for the weather focus area, with the following acronym descriptions.

- ASOS: Automated Surface Observing Systems

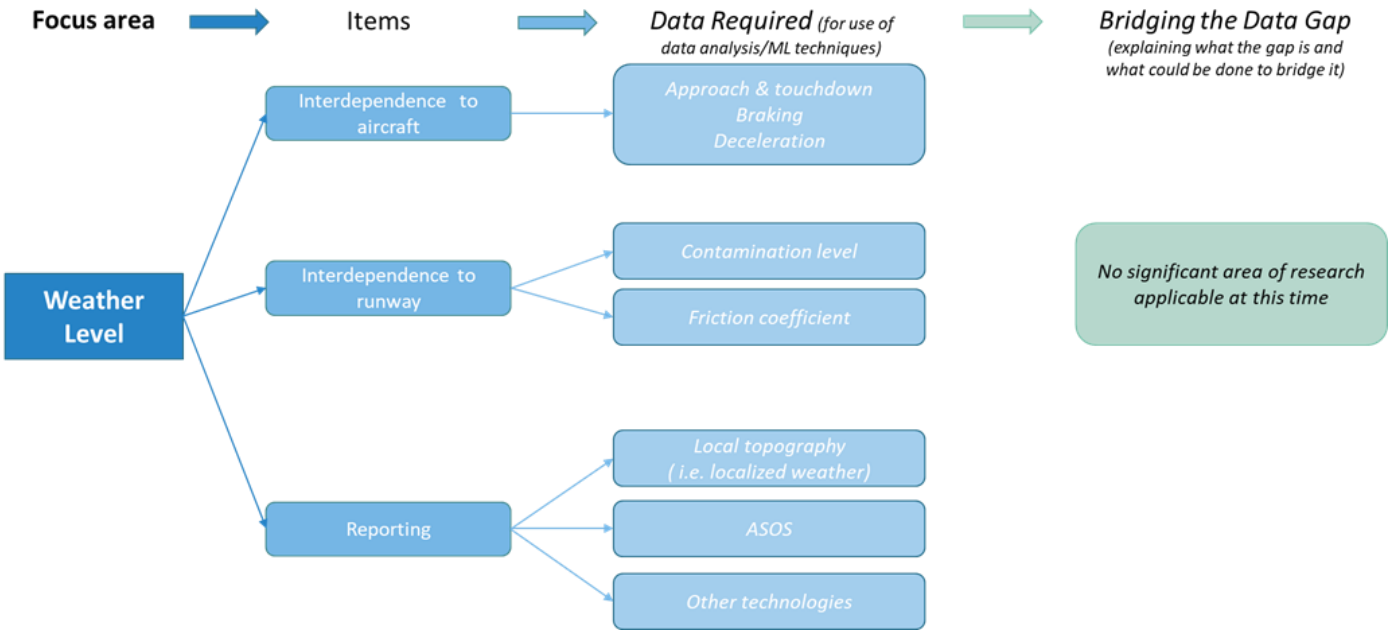


Figure 64. Weather focus area roadmap

10 References

Abadal, S., Jain, A., Guirado, R., López-Alonso, J., & Alarcón, E. (2021). Computing graph neural networks: A survey from algorithms to accelerators. *ACM Computing Surveys (CSUR)*, 1-38.

Ackley, J. L. (2020). A Supervised Learning Approach for Safety Event Precursor Identification in Commercial Aviation. *AIAA AVIATION 2020 FORUM*.

AlexKlein-Paste. (2015). A decision support model to assess the braking performance on snow and ice contaminated runways. *Cold Regions Science and Technology, 117*, 43-51.

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*, 272-278.

ASTM. (2020, November 11). *Standard Guide for Friction-Limited Aircraft Braking Measurements and Reporting.* Retrieved from ASTM International: https://www.astm.org/e3266-20.html

Belete, D. M., & Huchaiah, M. D. (2021). Grid search in hyperparameter optimization of machine learning models for prediction of HIV-AIDS test results. *International Journal of Computers and Applications*, 1-12.

Campbell, A. (2016). Uncertainty Analysis for Calculating Reverse Thrust using In Situ Data. *16th AIAA Aviation Technology, Integration, and Operations Conference.* Washington, D.C.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *22nd ACM SIGKDD international conference on knowledge discovery and data mining*, (pp. 785-794).

Chen, T., He, T., Benesty, M., & Khotilovich, V. (2019). *Package 'xgboost'*. Retrieved April 6, 2022, from https://cran.r-project.org/web/packages/xgboost/index.html

Daidzic, N. E. (2017). Modeling and Computation of the Maximum Braking Energy Speed for Transport Category Airplanes. *Journal of Aviation T4echnology and Engineering, 6*(2).

EASA. (2010). *Runway Friction Characteristics Measurement and Aircraft Braking (RuFAB).* European Aviation Safety Agency , Kanata.

EASA. (2017). *Review of Accident Presursors for Runway Excursions.* Cologne, Germany: European Aviation Safety Agency.

Es, G. V. (2001). *Hydroplanning of modern aircraft tires.* National Aerospace Laboratory NLR.

Es, G. V. (2017). Braking Capabilities on Flooded Runways: Flight Test Results Obtained with a Business Jet. *AIAA.* Amsterdam, The Netherlands.

ESDU. (2017). *Frictional and Retarding Forces on Aircraft Tyres.* United Kingdom: Engineering Sciences Data Unit (ESDU).

FAA. (2015). *Advisory Circular 25-31.* US Department of Transportation.

FAA. (2018, February 20). *AC 91-79A - Mitigating the Risks of a Runway Overrun Upon Landing.* Retrieved from Federal Aviation Administration: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_91-79A_Chg_2.pdf

FAA. (2020, October 29). *AC 150/5200-30D: Airport Field Condition Assessments and Winter Operations Safety.* Retrieved from https://www.faa.gov/documentLibrary/media/Advisory_Circular/150-5200-30D-chg-2-consolidated.pdf

FSF. (2009). *Reducing the Risk of Runway Excursions.* Flight Safety Foundation.

Hanyu, G., Struble, T. J., Coley, C. W., Wang, Y., Green, W. H., & Jensen, K. F. (2018). Using machine learning to predict suitable conditions for orgnaic reactions. *ACS central science*, 1465-1476.

IGGA. (2009, September). *Grooved runway surfaces drastically reduce all types of skids, including dynamic hydroplanning.* Retrieved April 6, 2022, from https://www.igga.net/wp-content/uploads/2018/08/FSSept2009_airport_grooving.pdf

Janakiraman, V. M. (2017). *Explaining Aviation Safety Incidents Using Deep Learned Precursors.* Moffett Field, CA 94035, USA: USRA/ NASA Ames Research Center.

Jarry, G. (2020). Approach and Landing Aircraft On-Board Parameters Estimation with LSTM Networks. *International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT).* Singapore.

Jasra, S. K.-M. (2018). Literature review of machine learning techniques to analyse flight data. *Computer Science*.

Jenkins, M. (2012). Reducing Runway Landing Overruns. Boeing AeroMagazine.

Kang, Z. (2020). A Deep Sequence-to-Sequence Method for Aircraft Landing Speed Prediction Based on QAR Data. *Web Information Systems Engineering – WISE 2020* .

Kirkland, I., Caves, R., Hirst, M., & Pitfield, D. (2003). The normalisation of aircraft overrun accident data. *Journal of Air Transportation Management, 9*(6), 333-341. doi:10.1016/S0969-6997(03)00033-4

Klein-Paste, A. (2012). Braking performance of commercial airplanes during operation on winter contaminated runways. *Cold Regions Science and Technology, 29*(37), 79-80.

Lee, H. (2020). Critical Parameter Identification for Safety Events in Commercial Aviation Using Machine Learning. *Aerospace, 6*(7), 73.

Li, L. (2015). Analysis of Flight Data Using Clustering Techniques for Detecting Abnormal Operations. *Journal of Aerospace Information Systems, 12*(9), 587-598.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 6765-6816.

Lv, H. (2018). A Novel Method of Overrun Risk Measurement and Assessment using Large Scale QAR Data. *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService).* Bamberg, Germany.

Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M., & Webster, S. (2000). An investigation of machine learning based prediction systems. *Journal of systems and software*, 23-29.

Mangortey, E. (2020). Application of Machine Learning Techniques to Parameter Selection for Flight Risk Identification. *AIAA Scitech 2020 Forum.* Orlanda, FL.

Mangortey, E., Monteiro, D., Ackley, J., Gao, Z., Puranik, T. G., Kirby, M., . . . Mavris, D. N. (2020). Application of Machine Learning Techniques to Parameter Selection for Flight Risk Identification. *AIAA SciTech 2020 Forum*, (p. 1850).

Memarzadeh, M. (2021). Multi-Class Anomaly Detection in Flight Data Using Semi-Supervised Explainable Deep Learning Model. *AIAA Scitech 2021 Forum.* AIAA Scitech 2021 Forum.

Nanduri, A. (2016). Anomaly detection in aircraft data using Recurrent Neural Networks (RNN). *2016 Integrated Communications Navigation and Surveillance (ICNS).* Herndon, VA, USA.

NTSB. (2016, October 6). *A-16-020-029.* Retrieved from National Transportation Safety Board: https://www.ntsb.gov/safety/safety-recs/RecLetters/A-16-020-029.pdf

OCallaghan, J. J. (2016). Slippery When Wet: The Case for More Conservative Wet Runway Braking Coefficient Models. *AIAA Technology, Integration, and Operations Conference*.

Odisho, E. (2020). Predicting Pilot Misperception of Runway Excursion Risk Through Machine Learning Algorithms of Recorded Flight Data. Embry-riddle Aeronautical University.

Pasindu, H. R. (2011). Computation of Aircraft Braking Distances. *Transportation Research Record: Journal of the Transportation Research Board, 2214*(1).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . al., e. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 2825-2830.

Puranik, T. (2020). Towards online prediction of safety-critical landing metrics in aviation using supervised machine learning. *Transportation Research Part C: Emerging Technologies, 120*.

Roginski, M. (2012). Manufacturer's PerspectiveRunway Friction and Aircraft Performance. Panama City: Boeing Airport Technology Group.

Sekhar, C. R., & Minal, E. M. (2016). Mode Choice Analysis Using Random Forest Decision Trees. *Transportation Research Procedia*, 644-652.

Sheridan, K., Puranik, T. G., Mangortey, E., Pinon-Fischer, O. J., Kirby, M., & Mavris, D. N. (2020). An application of DBSCAN clustering for flight anomaly detection during the approach phase. *AIAA SciTech 2020*, (p. 1851).

Stahle, L., & Wold, S. (1989). Analysis of Variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems*, 259-272.

Tong, C. (2018). An innovative deep architecture for aircraft hard landing prediction based on time-series sensor data. *Applied Soft Computing, 73*, 344-349.

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. *International work-conference on artificial neural networks*, (pp. 758-770).

Wahi, M. K. (2012). Airplane Brake-Energy Analysis and Stopping Performance Simulation. *Journal of Aircraft, 10*, 688.

Yacouby, R., & Axman, D. (2020). Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classificaiton Models. *First Workshop on Evaluation and Comparison*, (pp. 79-91).

Yager, T. J. (2013). *Factors influencing aircraft ground handling performance.* NASA Langley Research Center Hampton, VA, United States: NASA.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access*, 21020-21031.

Zhang, H. (2018). Aircraft Hard Landing Prediction Using LSTM Neural Network. *2nd International Symposium on Computer Science and Intelligent Control.*

A Remaining airframes clustering results

In the following figures, the graphs to the left illustrate the clustering results of each airframe flight data by varying the epsilon parameter from 0.1 to 1, while the graphs to the right illustrate the clustering results (using the DBSCAN algorithm) with an epsilon parameter value corresponding to when the outlier flights represent less than 5% of the total flights. This is because all of the flights landed safely so the DBSCAN is set up to separate the individual flights that are unique from the rest of the flights. The numbers in the legend show the flight counts per cluster. The DBSCAN is trained with single-point and calculated parameters. The training data is normalized per parameter so that the magnitude of each parameter does not contribute to the clustering.

NB-B2/2 Airframe

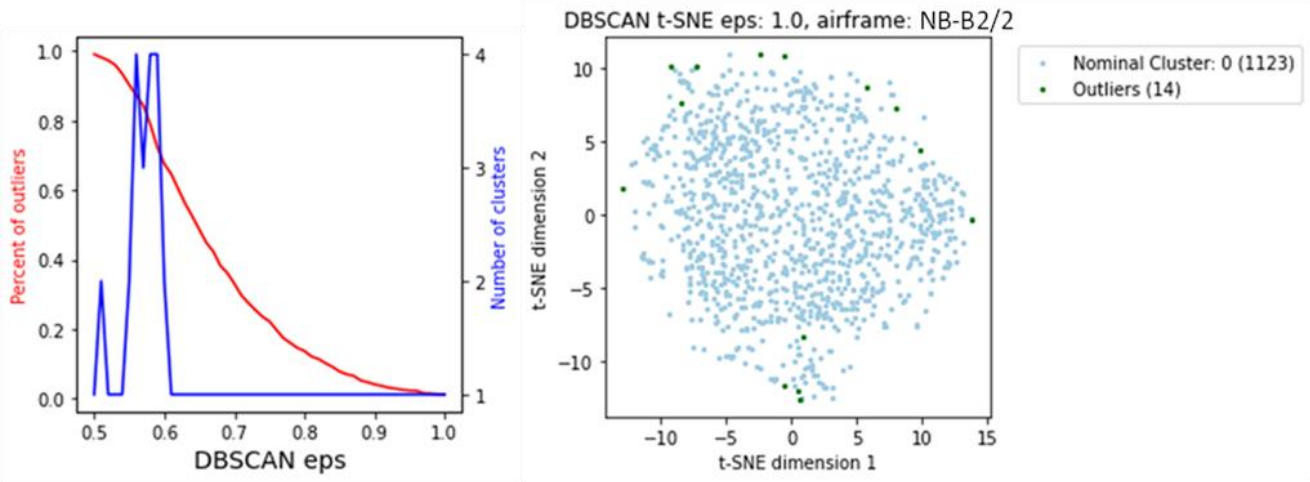


Figure A-1. DBSCAN cluster results for the NB-B2/2 airframe

NB-B2/3 Airframe

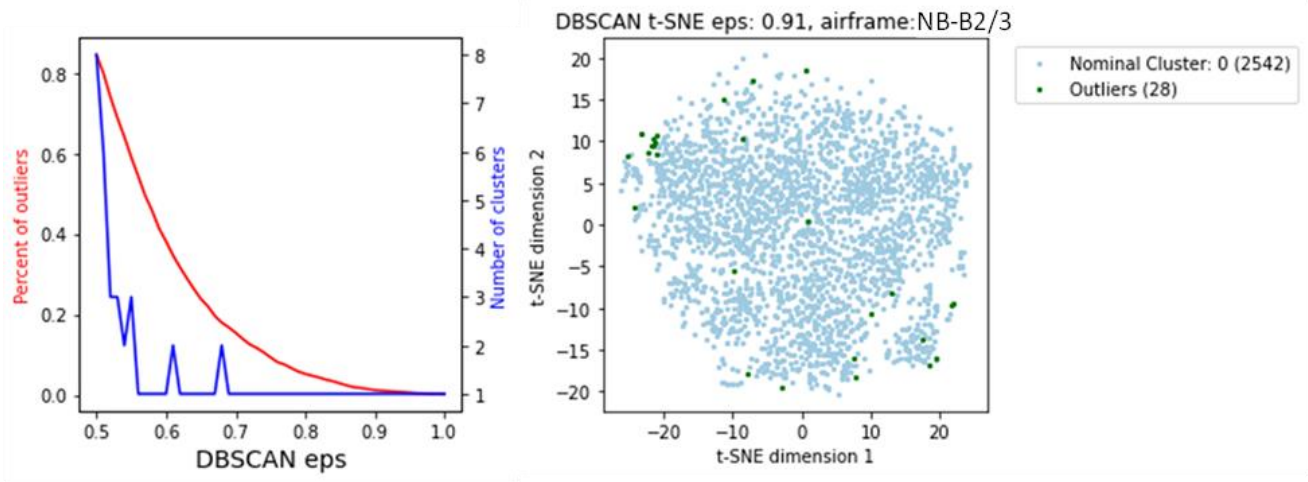


Figure A-2. DBSCAN cluster results for the NB-B2/3 airframe

NB-A2 Airframe

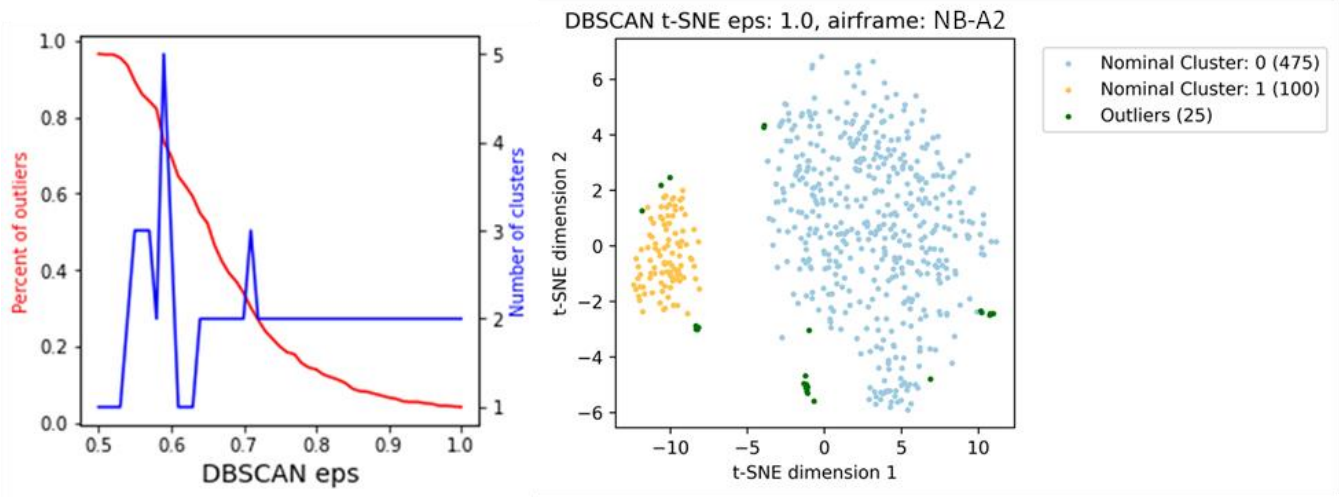


Figure A-3. DBSCAN cluster results for the NB-A2 airframe

WB-A1/2 Airframe

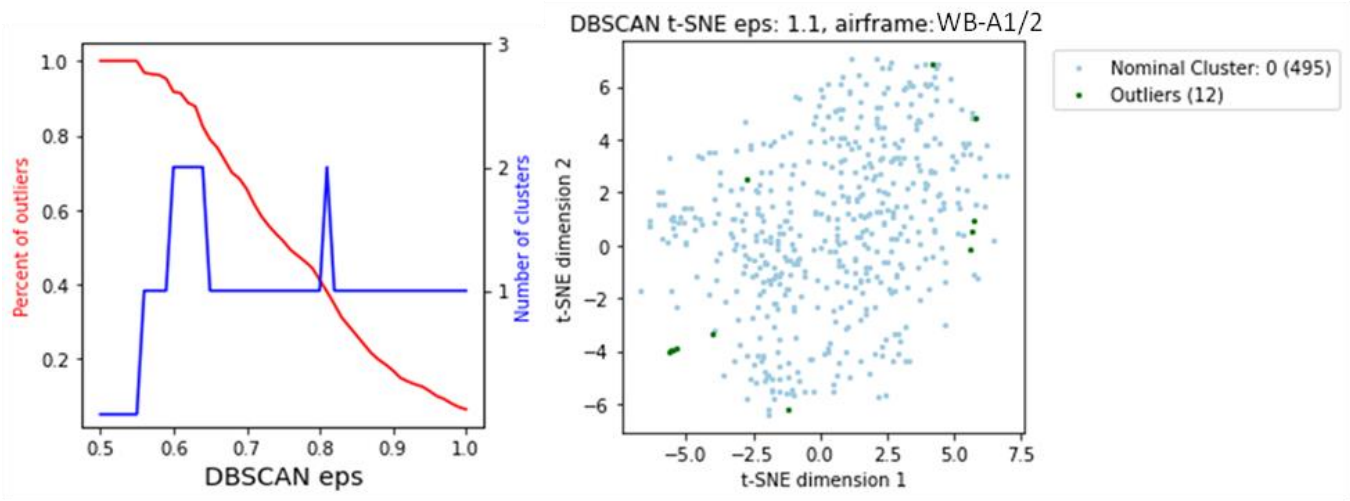


Figure A-4. DBSCAN cluster results for the WB-A1/2 airframe

NB-C1 Airframe

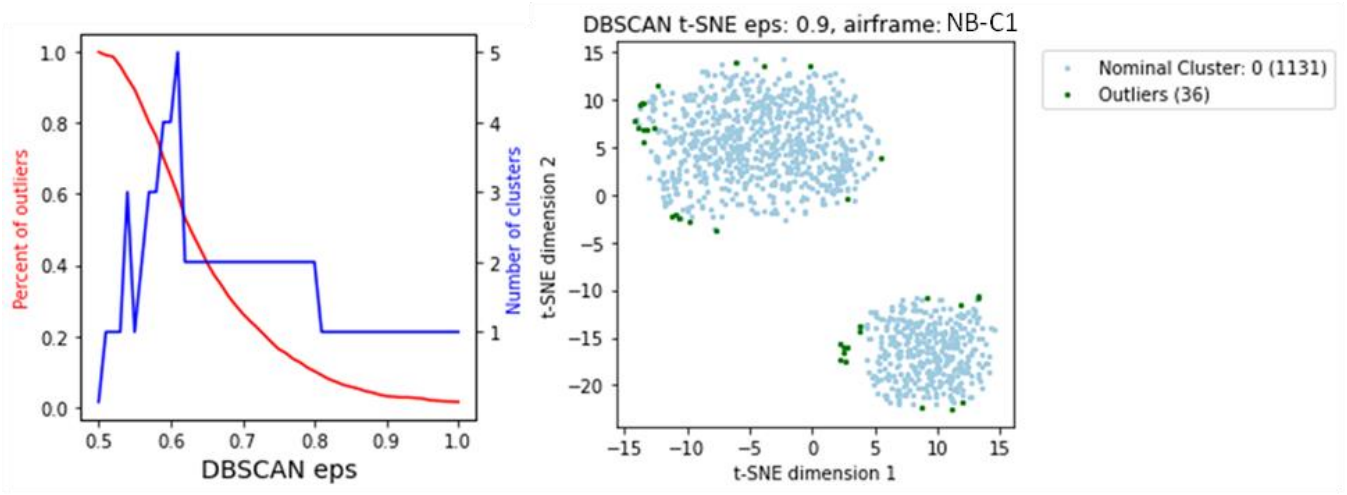


Figure A-5. DBSCAN cluster results for the NB-C1 airframe

B Remaining airframes decision tree results

The results below show the decision trees for each airframe generated by fitting the cluster label to the flight data. Each leaf of the decision tree describes the parameter name, decision threshold magnitude, “Gini” value (cleanness of split), number of samples/flights, number of samples/flights after the split, and the majority cluster representing the leaf. A feature of important plot is generated for each decision tree based on the “Gini” contribution of each parameter. Finally, the confusion matrix that shows the accuracy of the decision tree for representing the data cluster is displayed. The higher the values in the top left and bottom right of the matrix, the higher the accuracy.

NB-B2/2 Airframe

Confusion Matrix:

$$\begin{bmatrix} 14 & 0 \\ 0 & 1123 \end{bmatrix}$$

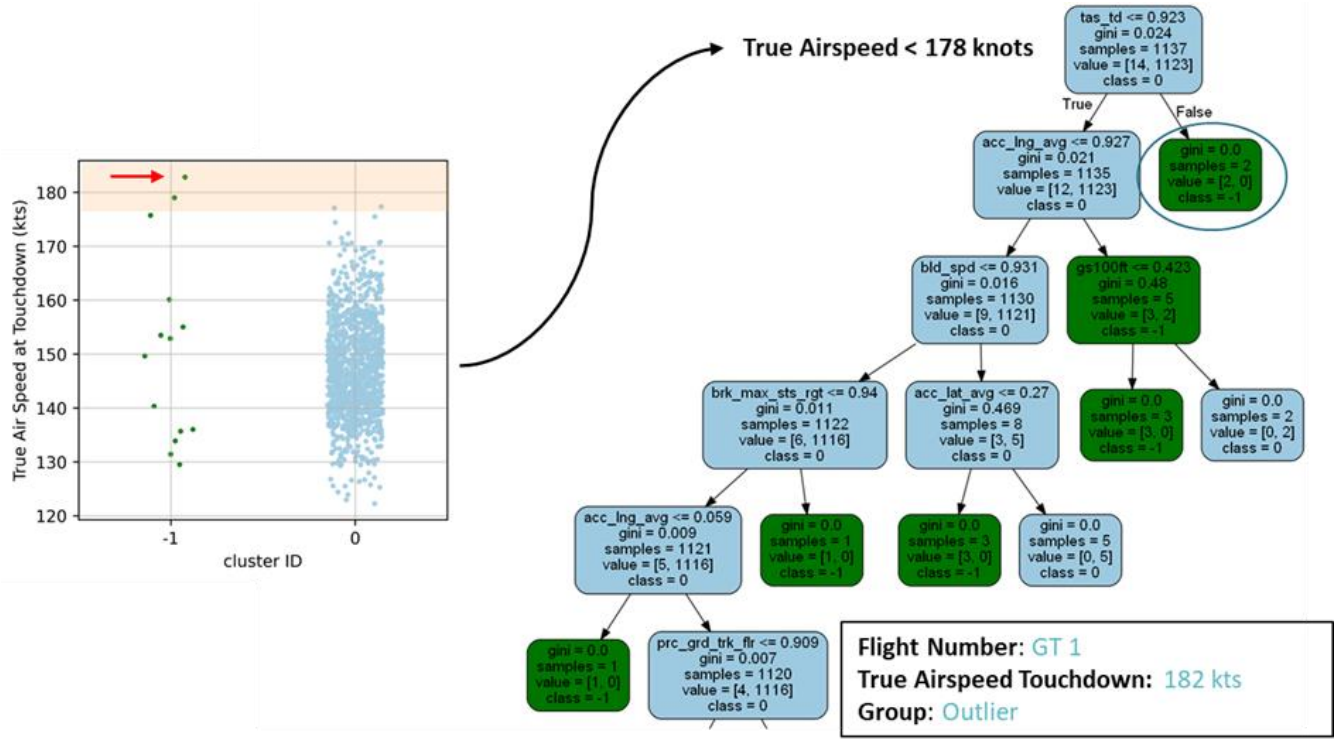


Figure B-1. Decision tree for the NB-B2/2 airframe trained with DBSCAN clusters and sample outlier identified at the upper level

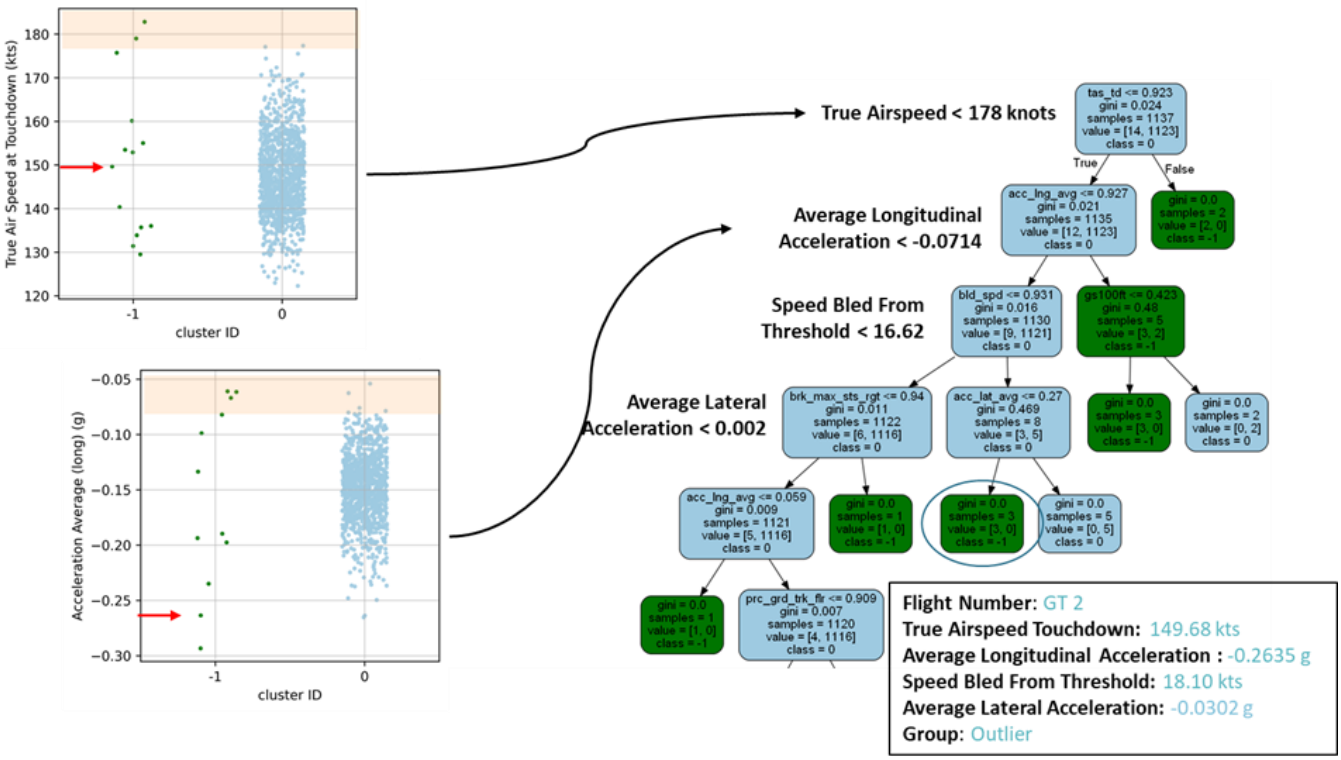


Figure B-2. Decision tree for the NB-B2/2 airframe trained with DBSCAN clusters and sample outlier identified at the mid-level

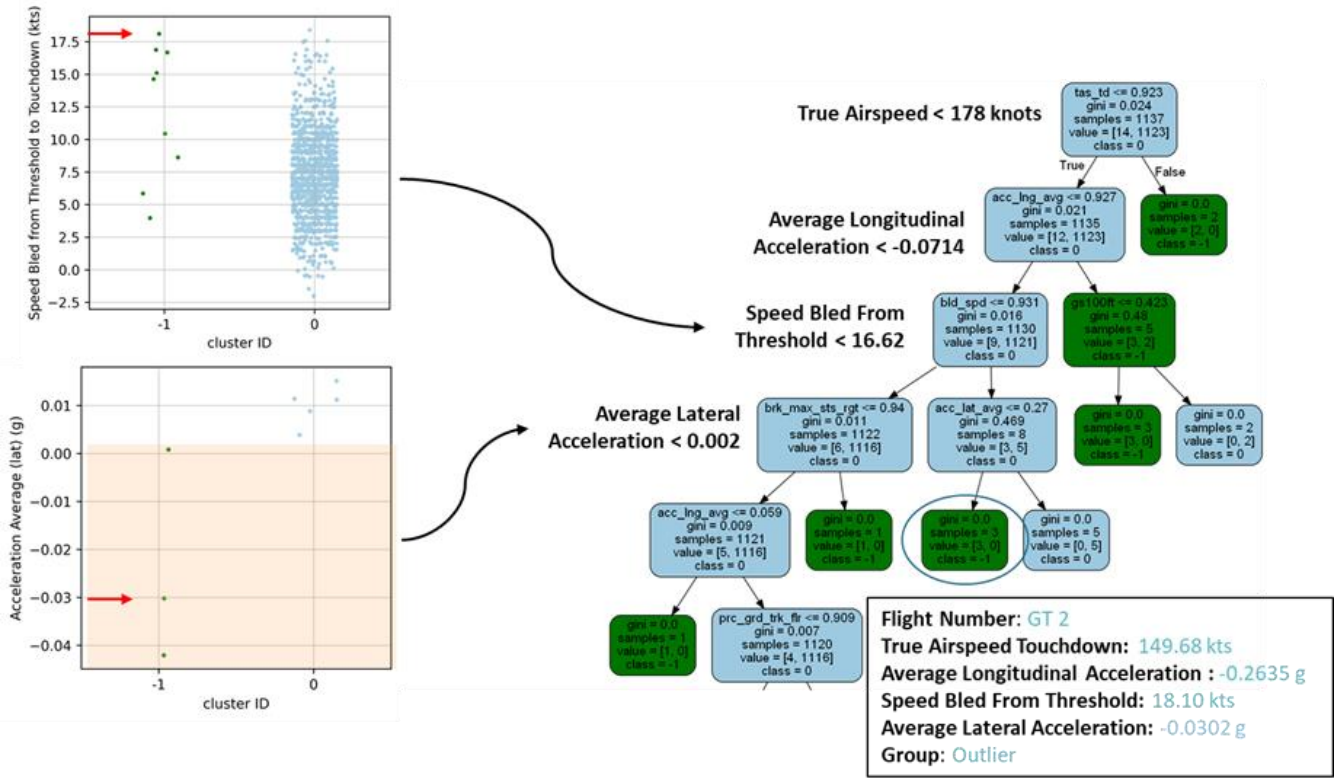


Figure B-3. Decision tree for the NB-B2/2 airframe with DBSCAN clusters and sample outlier identified at the lower level

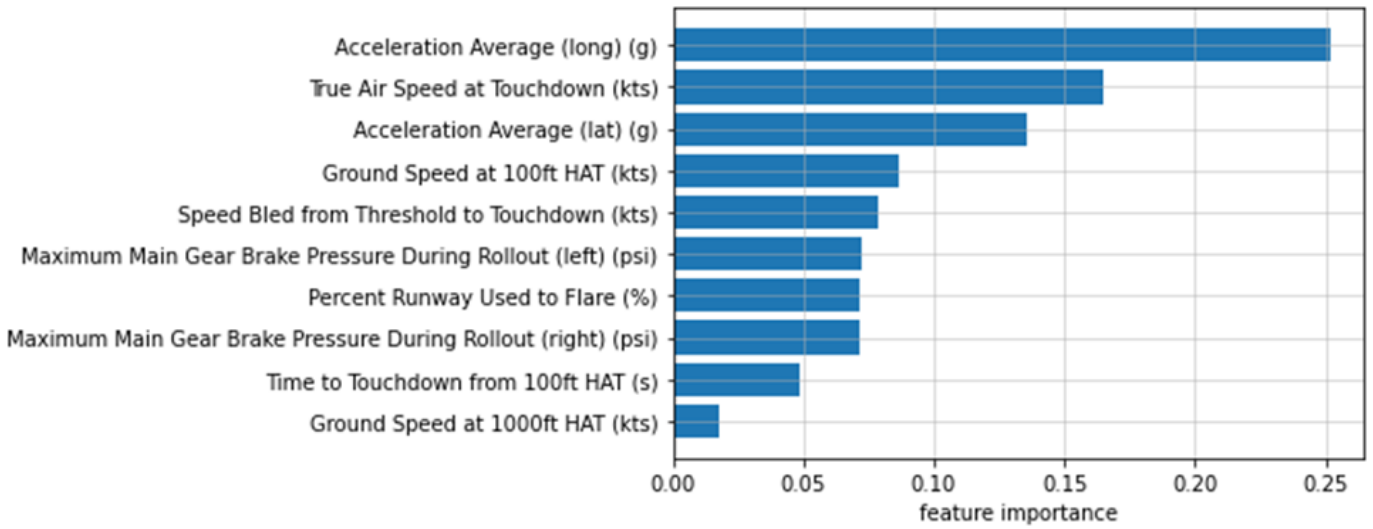


Figure B-4. Feature importance for the NB-B2/2 airframe decision tree

NB-B2/3 Airframe

Confusion Matrix:

$$\begin{bmatrix} 20 & 8 \\ 0 & 2542 \end{bmatrix}$$

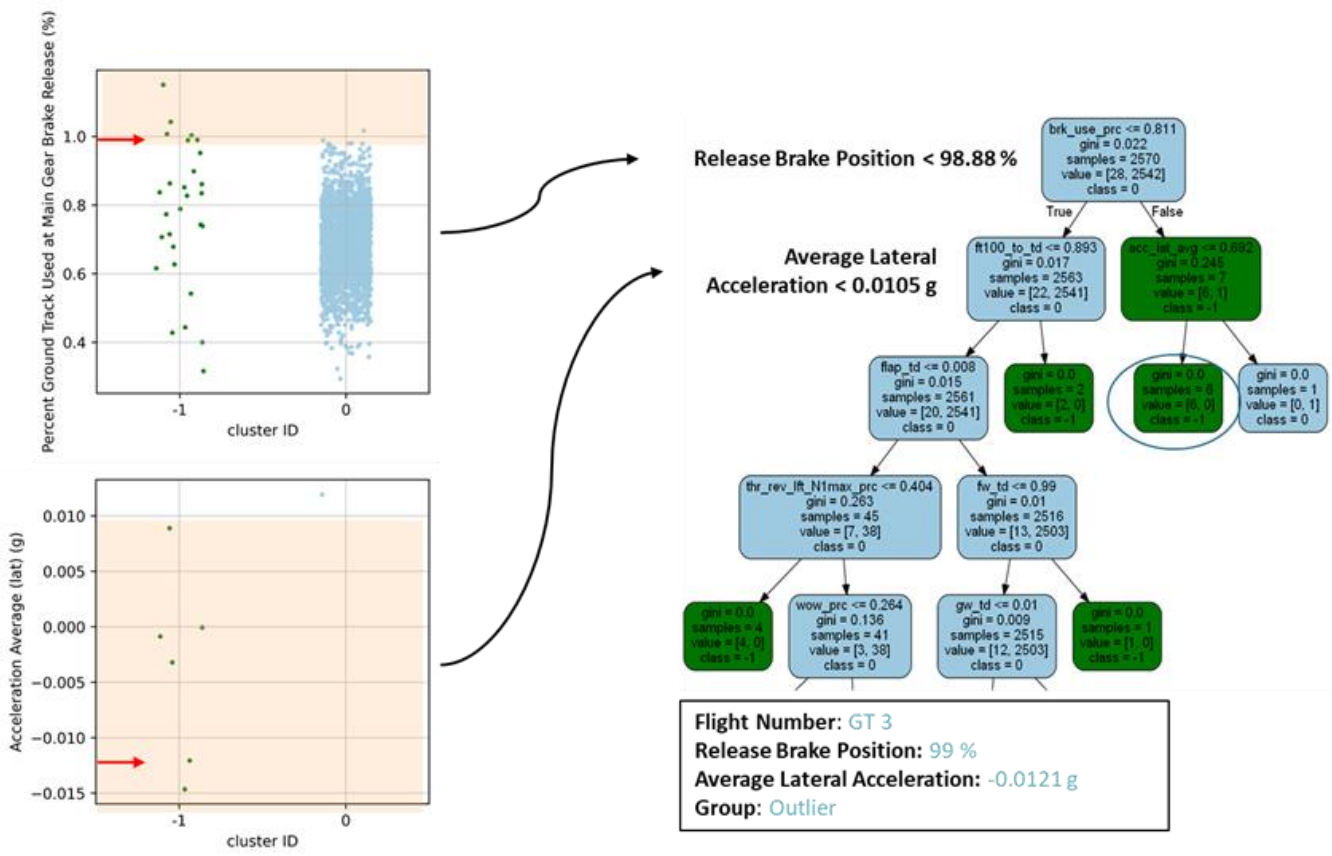


Figure B-5. Decision tree for the NB-B2/3 airframe with DBSCAN clusters and sample outlier identified at the upper level

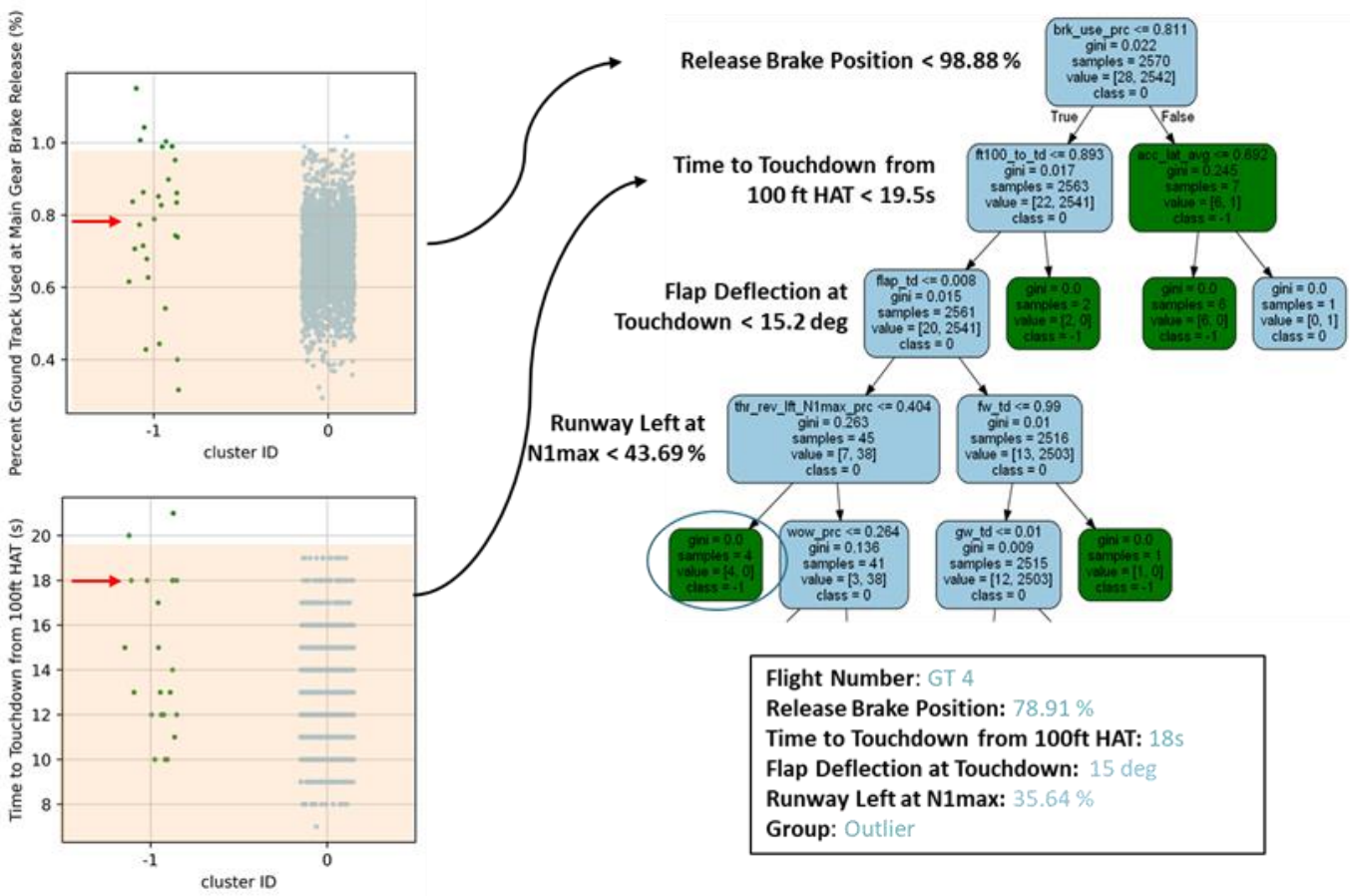


Figure B-6. Decision tree for the NB-B2/3 airframe with DBSCAN clusters and sample outlier identified at the mid-level

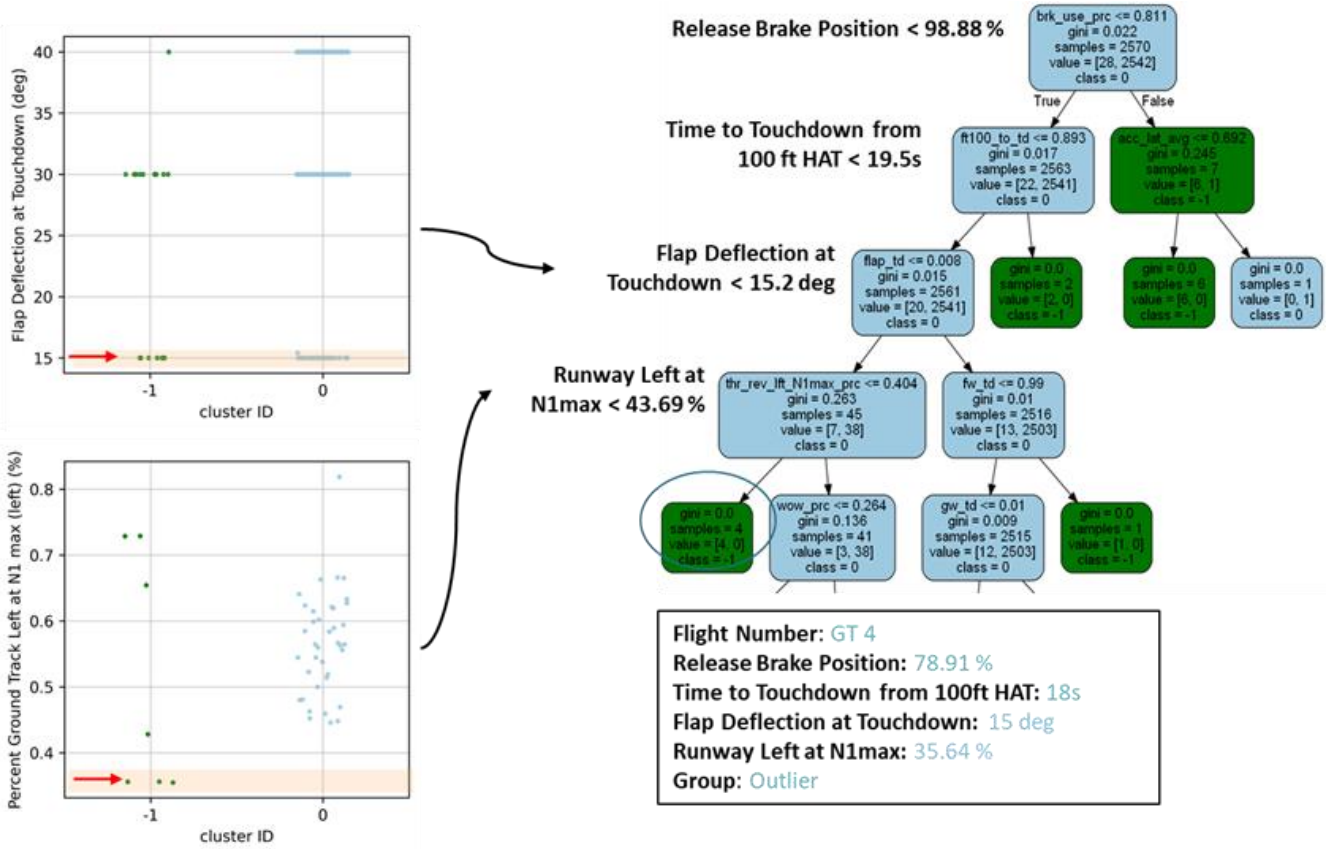


Figure B-7. Decision tree for the NB-B2/3 airframe with DBSCAN clusters and sample outlier identified at the lower level

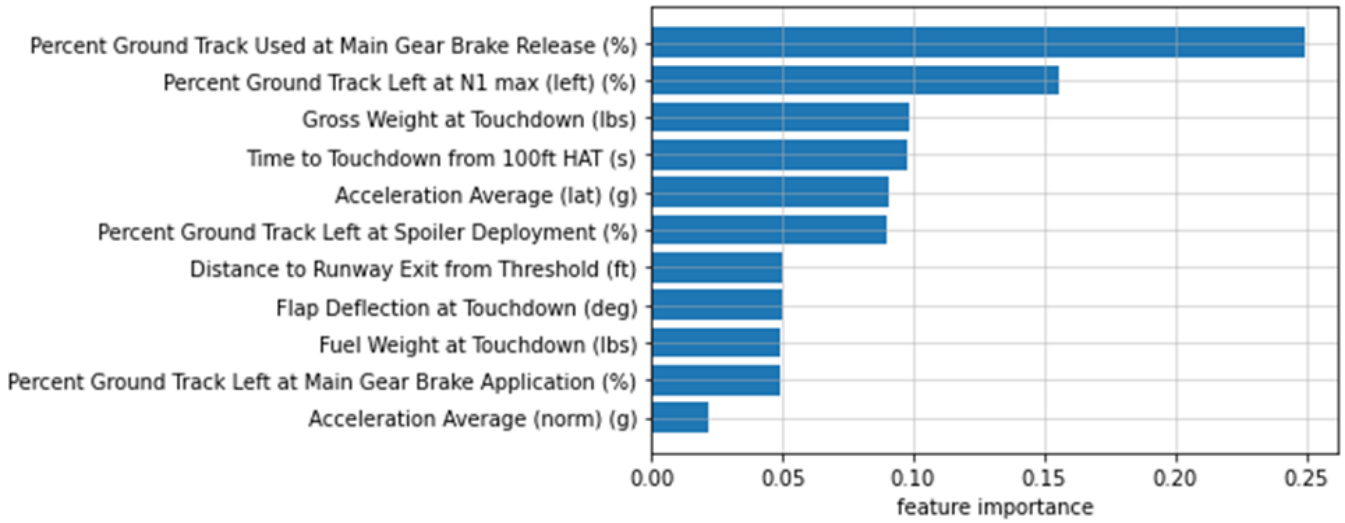


Figure B-8. Feature importance for the NB-B2/3 airframe decision tree

NB-A2 Airframe

Confusion Matrix:

$$\begin{bmatrix} 25 & 0 \\ 0 & 575 \end{bmatrix}$$

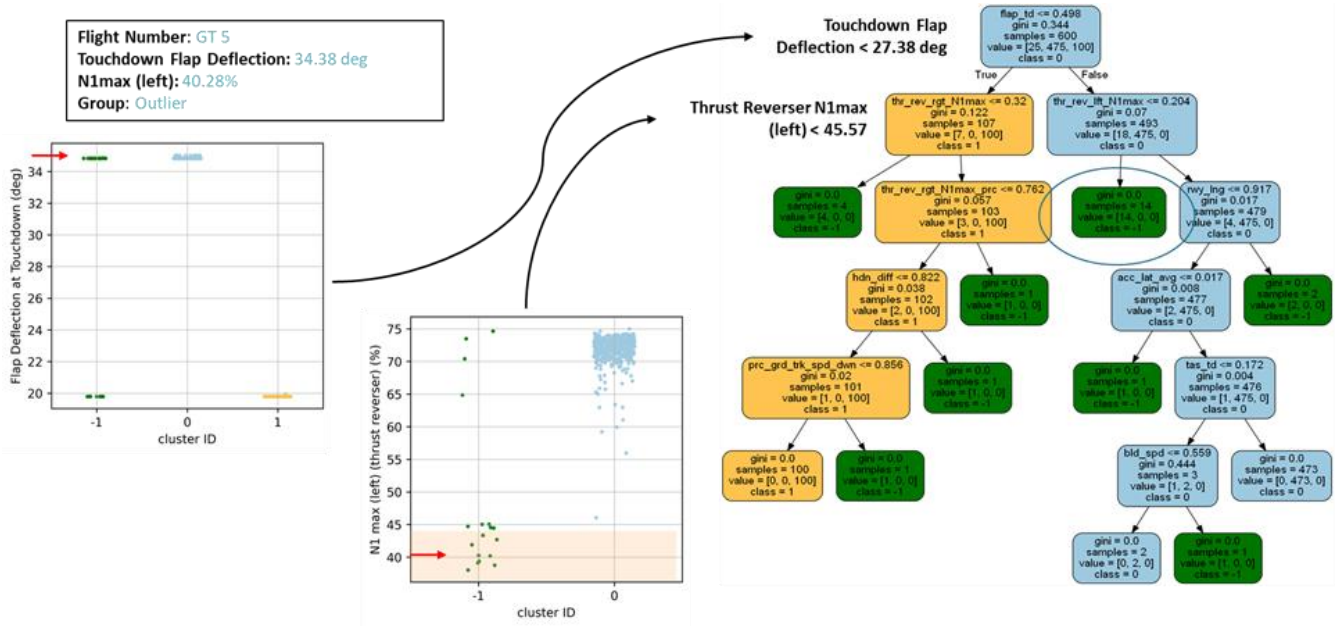


Figure B-9. Decision tree for the NB-A2 airframe with DBSCAN clusters and sample outlier identified at the upper level

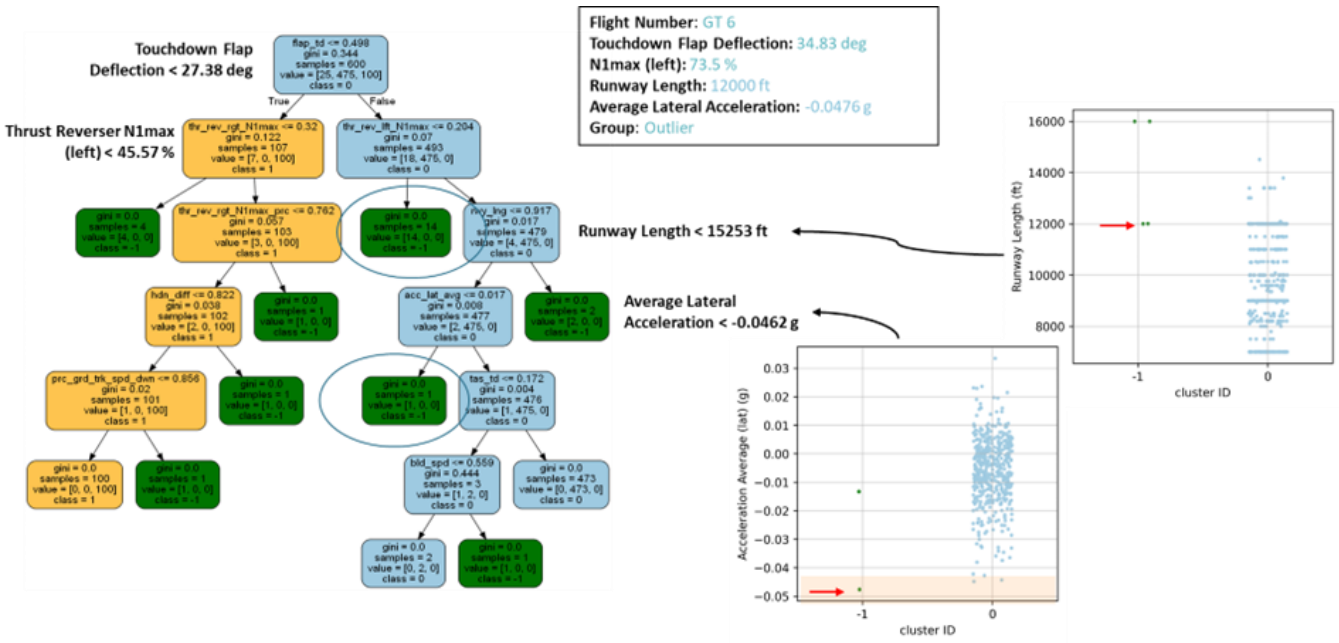


Figure B-10. Decision tree for the NB-A2 airframe with DBSCAN clusters and sample outlier identified at the mid-level

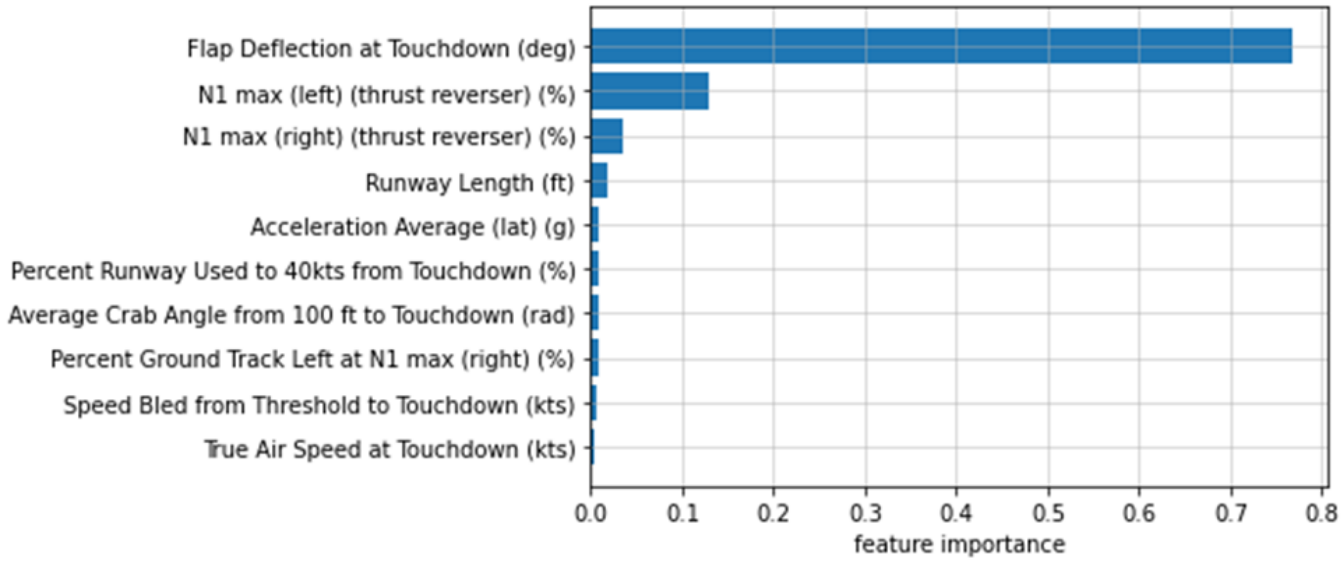


Figure B-11. Feature importance for the NB-A2 airframe decision tree

WB-A1/2 Airframe

Confusion Matrix

$$\begin{bmatrix} 31 & 5 \\ 0 & 1131 \end{bmatrix}$$

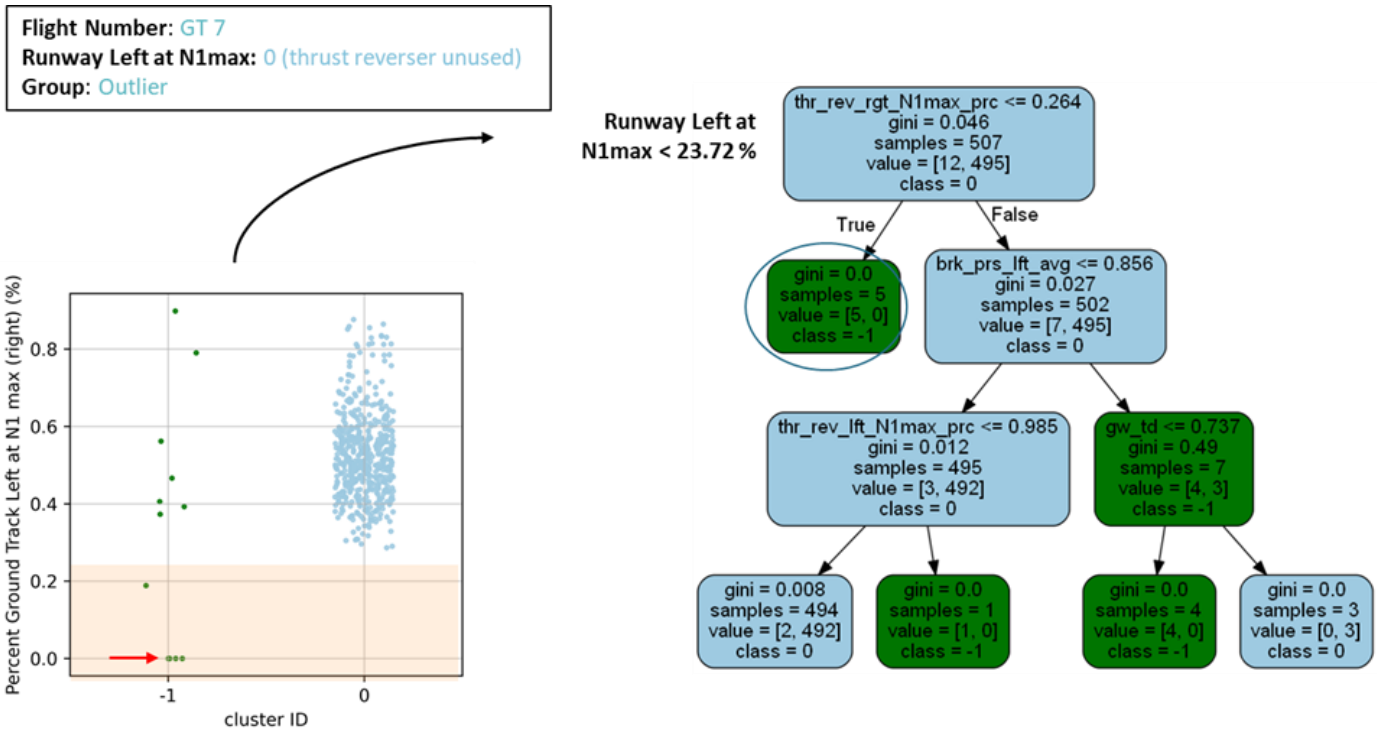


Figure B-12. Decision tree for WB-A1/2 airframe with DBSCAN clusters and sample outlier identified at the upper level

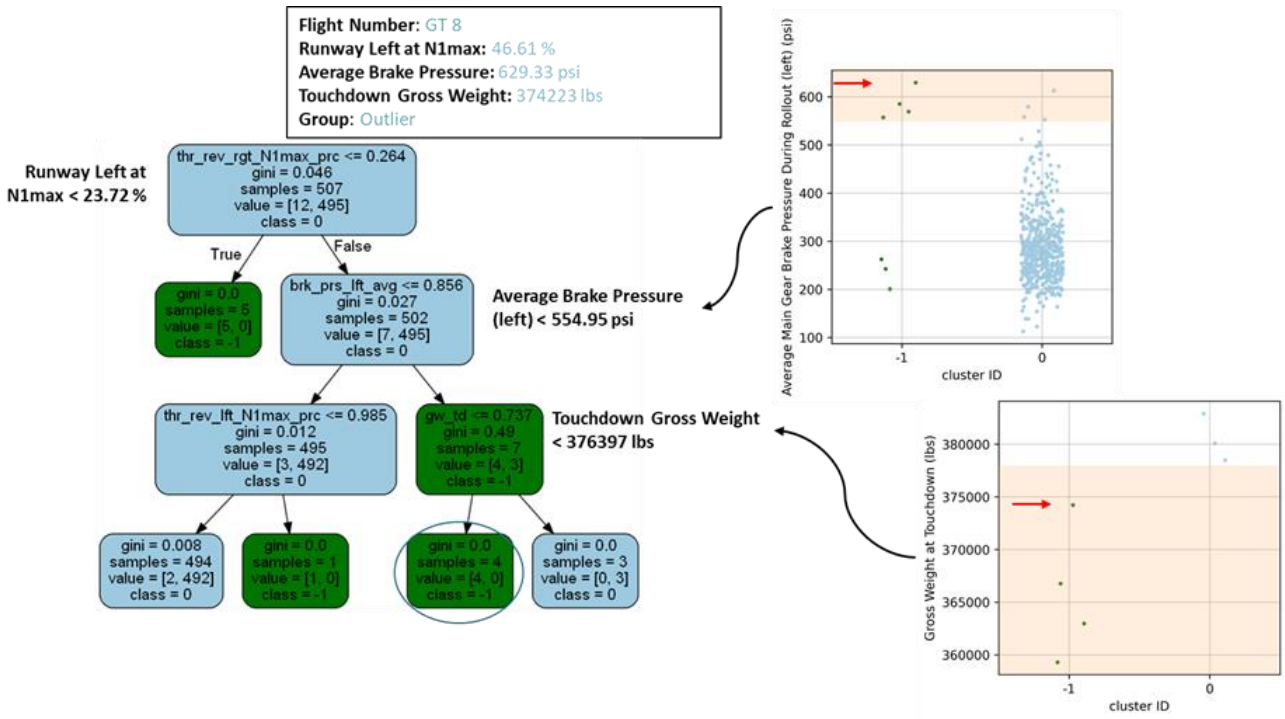


Figure B-13. Decision tree for the WB-A1/2 airframe with DBSCAN clusters and sample outlier identified at the lower level

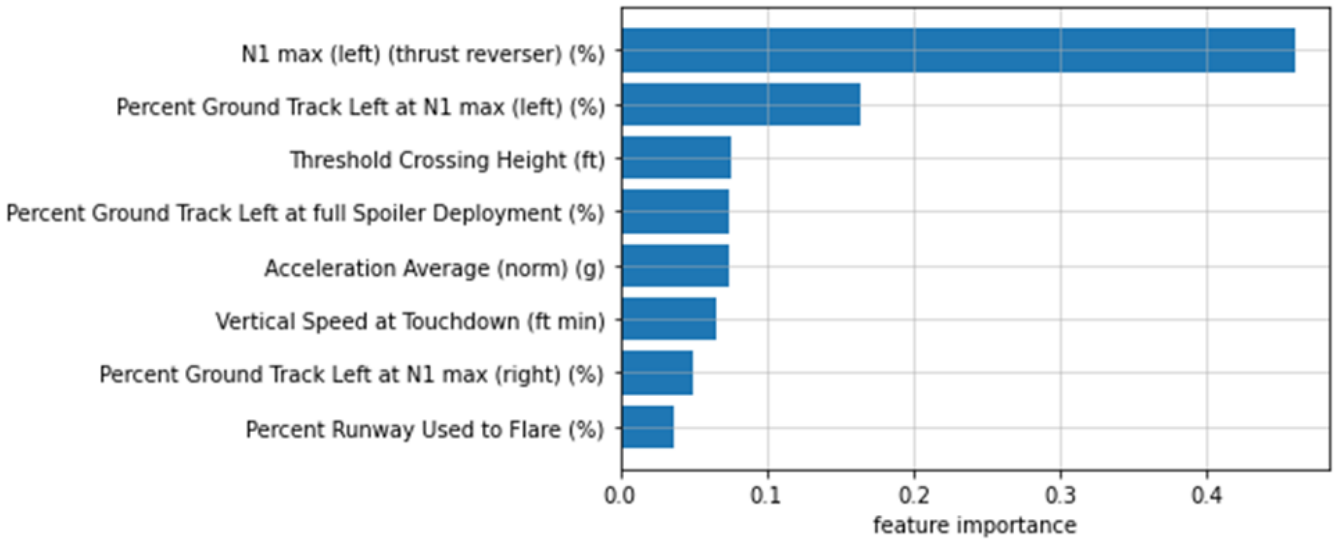


Figure B-14. Feature importance for the WB-A1/2 airframe decision tree

NB-C1 Airframe

Confusion Matrix

$$\begin{bmatrix} 31 & 5 \\ 0 & 1131 \end{bmatrix}$$

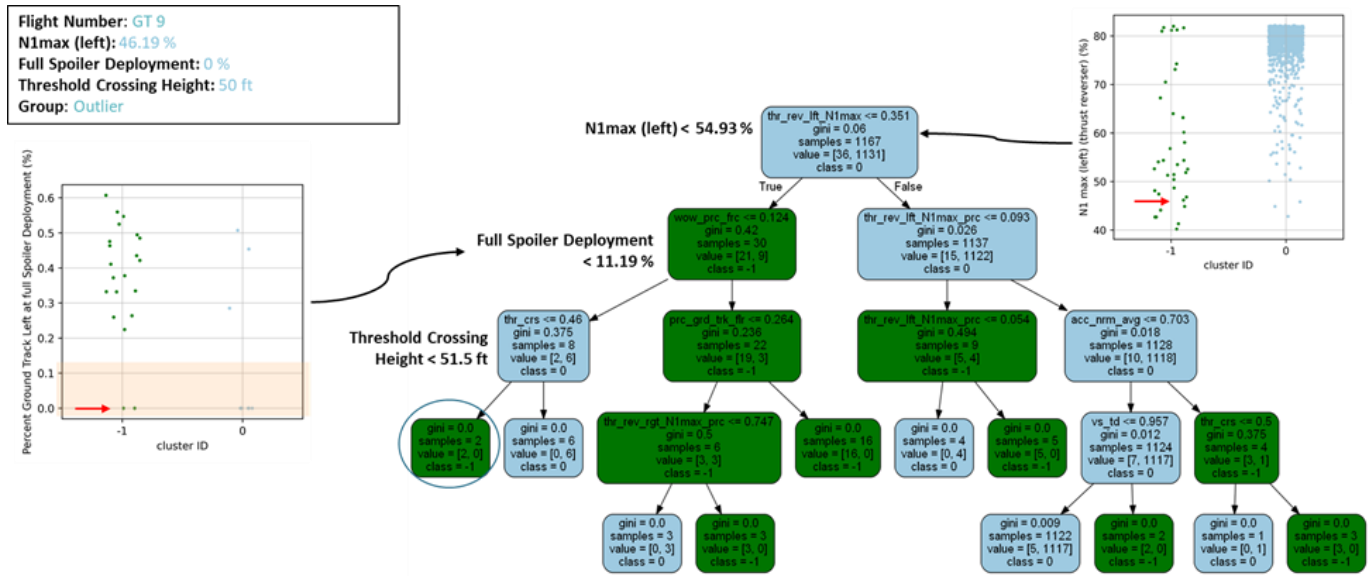


Figure B-15. Decision tree for the NB-C1 airframe with DBSCAN clusters and sample outlier identified under first branch

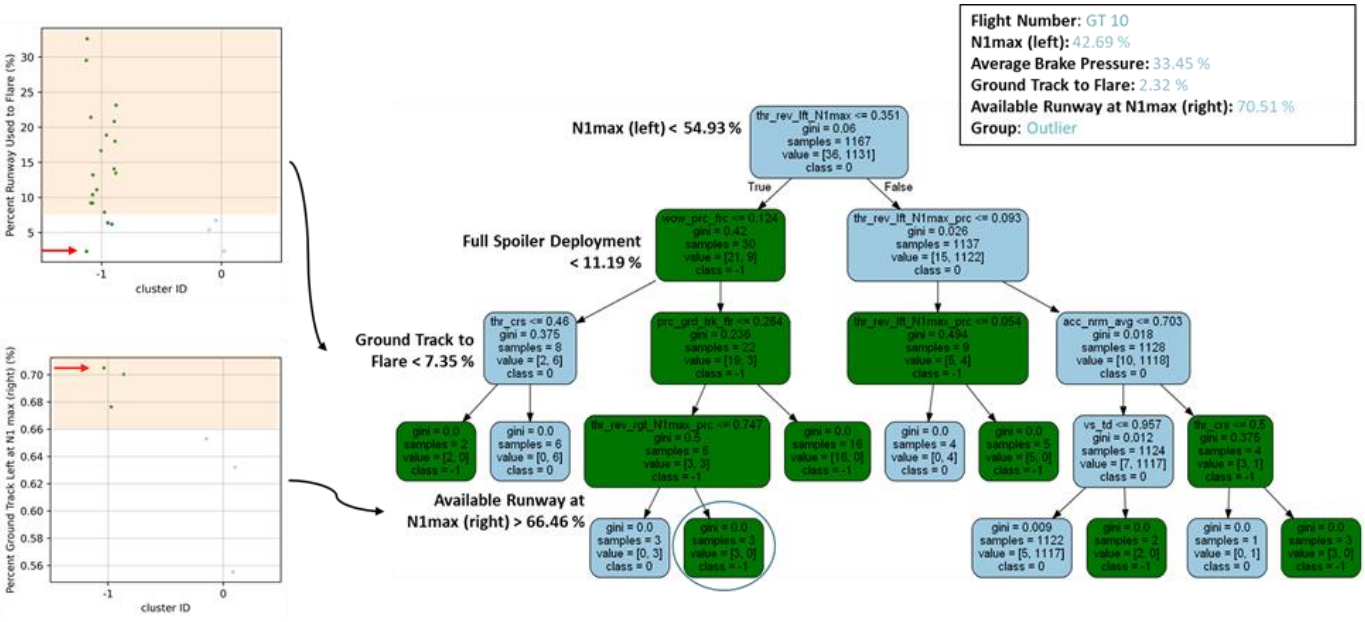


Figure B-16. Decision tree for the NB-C1 airframe with DBSCAN clusters and sample outlier identified under second branch

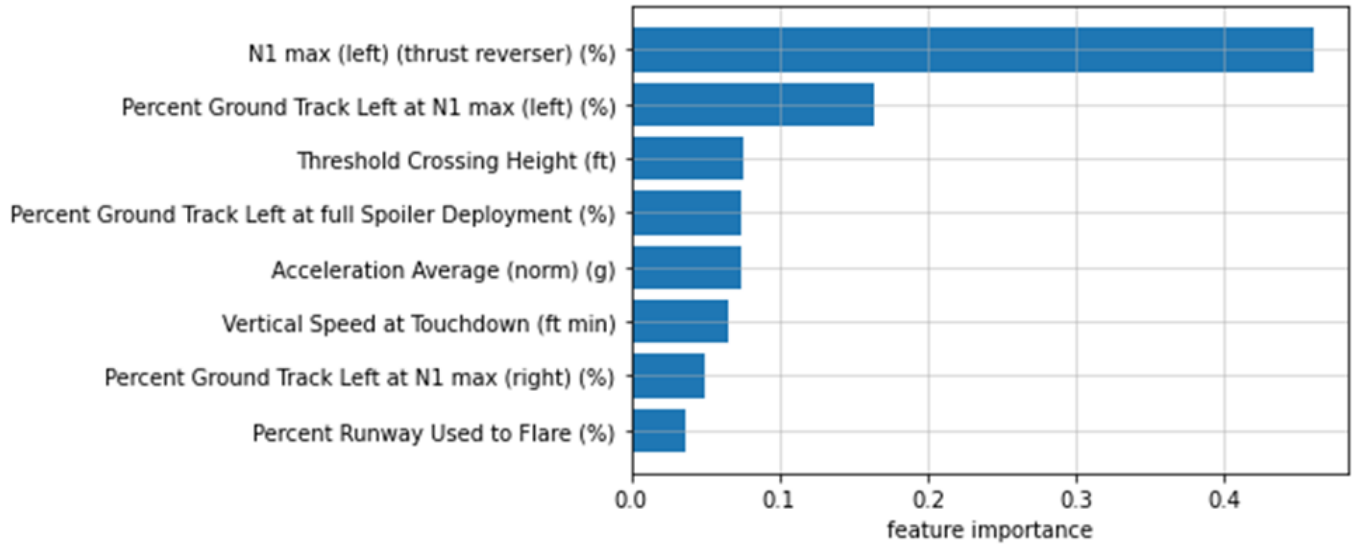


Figure B-17. Feature importance for the NB-C1 airframe decision tree