Predictive Modeling of High-Altitude Clear Air Turbulence in the United States: A Machine Learning Approach

Kadir Gökdeniz[1] · İrem Ülkü[2]

Abstract

High-altitude Clear Air Turbulence (CAT) poses significant risks to aviation safety due to its unpredictability and challenges in detection. This study leverages machine learning models to improve CAT prediction within U.S. airspace at 200–350 hPa pressure levels, utilizing Pilot Reports (PIREPs), ERA5 reanalysis data, and aircraft aerodynamic parameters from the BADA database. Gradient boosting algorithms, particularly XGBoost, achieved the highest performance with an AUC of 0.904, demonstrating superior capability in capturing nonlinear atmospheric dynamics. Key findings highlight the dominance of geographic coordinates (17.5% feature importance) and turbulence indices like TI3 in prediction, emphasizing the role of regional topography and upper-tropospheric instability. The integration of aerodynamic features, such as drag force and wing loading, enhanced the detection of moderate-to-severe turbulence (POD improved from 0.845 to 0.866), offering complementary value to traditional aircraft-agnostic methods. Seasonal analysis revealed winter months as peak periods for CAT incidents, correlating with jet stream activity. While results align with global studies, limitations include geographic scope and aircraft-type diversity. This research underscores the potential of machine learning for operational CAT forecasting, with recommendations for future work focusing on global data integration and real-time telemetry to address climate-driven turbulence trends.

Keywords: Clear Air Turbulence (CAT), Machine Learning, Aviation Safety, Gradient Boosting, Aerodynamic Features

Introduction

The safety and efficiency of aviation operations are directly influenced by meteorological conditions, which pose significant operational risks. Factors such as turbulence, low visibility, low-level wind shear, precipitation, icing, low cloud ceilings, volcanic ash, and sandstorms can disrupt flight schedules and pose risks to both human and aircraft safety.

Turbulence stands as one of the most critical factors affecting aviation safety, presenting significant risks to both passengers and aircraft. A study by Gultepe et al. reveals that 71% of aviation accidents caused by atmospheric conditions are attributed to turbulence (Gultepe et al. 2019). Moreover, turbulence is estimated to impose an annual economic cost of $100 to $200 million on the United States (Paul D. et al. 2017). However, turbulence is one of the most complex challenges in fluid dynamics due to its inherently chaotic and unpredictable nature. While the Navier-Stokes equations provide a solid theoretical foundation, solving these equations requires substantial computational power, limiting their practical applications (Wang Rui et al. 2020). Consequently, numerical weather predictions and various ensemble methods have been developed, but the need for innovative approaches remains.

Turbulence can be classified into several subtypes, each with unique dynamics. For instance, mountain wave turbulence occurs when wind crosses mountain ridges, creating a wavy airflow that poses significant risks to low-altitude flights (Sharman R. 2016). Convective turbulence, on the other hand, typically arises on hot days with intense vertical air movement or near thunderstorm clouds (Nilsson et al. 2001).

Clear Air Turbulence (CAT) is a phenomenon that occurs unexpectedly in clear skies and is challenging to observe directly by pilots or detect via radar systems (Venkatesh et al. 2013). This makes CAT one of the most unpredictable types of turbulence, granting it a special focus in aviation research. Notably, recent research suggests a rising trend in clear air turbulence across regions including East Asia, the Eastern Pacific, and the Northwestern Pacific. (Lee et al. 2022). Furthermore, a strong correlation has been highlighted between global warming and clear air turbulence, emphasizing the role of climate-induced atmospheric instabilities in this increase (Foudad 2024). Another study by Meneguz et al. shows that airflow over mountain ranges in the upper troposphere can elevate the likelihood of encountering CAT and orographic turbulence to over 80% (Meneguz E. et al. 2016).

Historically, numerical weather prediction (NWP) models have been the primary approach for analyzing clear air turbulence (CAT). (Kim et al. 2021, Kristoffer 2013). In recent years, numerical weather prediction has been integrated with machine learning algorithms

Table 1 Literature Review Summary

Reference	Region	Methods	Time Interval	Dataset	Time Resolution	Types of Turbulence	Results
Muñoz-Esparza et al. (2020)	USA	Random Forest, GBRT	June 2018-September 2019	[1] and [7]	6 hours	CAT, MWT	AUC: 0.770 - 0.906
Hon et al. (2020)	Asia-Pacific	XGBoost	April 2018 - March 2019	[1] and [2]	1 hour	MOG CAT	AUC score: 0.84
Shao et al. (2024)	China	XGBoost, SVM, LR, RF, DT	Jan 2020 - Dec 2022	[1],[2] and ERA5	1 hour	MOG CAT ,CIT	AUC scores: 0.75 - 0.96
Nerushev et al. (2022)	Northern Hemisphere	Correlation-extreme algorithms	2007-2018	[3],[4] and [5]	15 minutes (image interval)	CAT, MWT	$R^2 = 0.68$ - 0.74
Sharman et al. (2016)	United States	Statistical Mapping of Turbulence	—	[2], GFS, [6] and ERA5	6 hours	CAT, MWT	AUC: 0.77 - 0.91
Pearson et al. (2017)	US National Airspace	Integrate [1], [2], [8], [9], [10]	Jul-Sep 2015, Jan-Mar 2016	[2], [8], EDR, WRF-RAP	15 minutes	Convective, CAT, MWT	AUC: 0.75 - 0.88
Storer et al. (2020)	Airspace of Europe and the UK	[11]	Sep 2016 - Aug 2017	Boeing 747 and 777 aircraft	24, 27, 30, and 33 hours	ST, MTW, and CIT	AUC: 0.84 - 0.88
de Mello (2024)	Brazil	Random Forest	Mar 2018 - Mar 2021	GFS, ERA-5, VRTG	—	CAT (3 different intensities)	POD:0.74-0.78, FAR: 0.26 -0.31

for CAT forecasting, demonstrating their effectiveness (Hon et al. 2020, Muñoz-Esparza et al. 2020). Studies in the literature exhibit significant variability in terms of problem definitions, geographical application areas, study periods, data sources, resolution levels, turbulence intensities, and findings. Table 1 categorizes and summarizes some studies on CAT analysis based on various criteria.

Muñoz-Esparza et al. (2020) conducted a study on the weaknesses of the GTG method and proposed a machine learning approach based on Random Forest instead. The study applied EDR transformations and utilized various turbulence indicators derived from HRRR forecast data. For 6-hour forecasts of turbulence types such as CAT and MWT, the results achieved an AUC score of 0.906.

A study by Hon et al. (2020) focuses on using the Multi Index Consensus (MIC) method for turbulence prediction in the Asia-Pacific region (Hon et al. 2020). The research processes over 16,000 pilot report (PIREP) data and 19 different turbulence diagnostic features using the XGBoost algorithm, demonstrating an improvement in prediction performance ranging from 3% to 17%. This study highlights the benefits of applying machine learning techniques to local-scale turbulence forecasting.

The study conducted by Shao et al. (2024) applied five different machine learning methods to predict Moderate or Greater (MOG) CAT turbulence over China (Shao et al. 2024). By utilizing in situ observations, PIREPs, and the ERA5 dataset, the research analyzed the temporal and spatial characteristics of turbulence through 15 distinct turbulence diagnostics.

This study conducted by Nerushevet al.2022 analyzes turbulence areas in the upper troposphere using satellite data from 2007-2018, investigating the relationship between turbulence and jet streams and climatic parameters (Nerushev et al.2022). The study identifies turbulence areas through water vapor inhomogeneities and correlation-extremal algorithms and uses regression models to explain the temporal variability of strong turbulence areas.

[1] In situ observation
[2] PIREP
[3] SEVIRI Radiometer Data
[4] NCEP/NCAR Reanalysis Data
[5] NOAA Arctic Sea Ice Data
[6] WRF-RAP
[7] HRRR Predictions
[8] METAR
[9] NTDA
[10] DCIT
[11] Multi-diagnostic multi-model ensemble

In their 2016 study, Sharman and Pearson developed a statistical mapping method for turbulence forecasting (Sharman et al.2016). This method converts various turbulence indicators derived from NWP model outputs into the ICAO-standard EDR. The study evaluated the model's accuracy by utilizing ROC curves for different turbulence categories. The results demonstrated that ensemble averaging (GTG) achieved higher AUC scores and outperformed individual indicators.

The study by Pearson and Sharman in 2017 is a continuation of their 2016 work (Pearson et al. 2017). In the initial research, the performance of the GTG community model was enhanced by integrating data sources such as DCIT, NTDA, PIREP, and METAR, leading to the creation of a new model called GTGN. It was highlighted that GTGN performs better in capturing convective turbulence, especially during the summer months.

Storer et al. 2020 investigated multi-diagnostic, multi-model ensemble forecasting for aviation turbulence, finding that combining multiple diagnostics improved skill over single-model ensembles (Storer et al. 2020). MOGREPS-G outperformed EPS51, and EPS12 slightly improved results. Smaller ensemble spreads within multi-diagnostic ensembles yielded better performance, with further study needed on optimal ensemble size.

The study conducted by de Mello 2024 applies machine learning algorithms using GFS, ERA-5, and VRTG data for Clear Air Turbulence (CAT) prediction, covering part of the Brazilian border. The Random Forest algorithm demonstrated the best performance in CAT prediction, achieving POD of 0.74 and FAR of 0.31 with GFS data, and POD of 0.78 and FAR of 0.26 with ERA-5 data.

In this study, the PIREP and ERA5 datasets were used to analyze data with a 1-hour time resolution from 2022 to 2024 within the United States. The study focused on CAT classification and examined upper altitude data at 200, 225, 250, 300, and 350 hPA levels. The performance of XGBoost, RandomForest, LightGBM, AdaBoost, and Logistic Regression models was compared.

This study differs from the existing literature primarily by examining how incorporating aircraft type and aerodynamic features from pilot reports into the model could potentially enhance classification performance. This approach offers an alternative to aircraft-type-independent classification methods such as the standard GTG, which are converted to ICAO-EDR.

Additionally, CAT was validated through pilot reports, and a human-centered classification system measuring the effects of turbulence on passenger comfort was investigated. A comprehensive evaluation of high-altitude CAT samples obtained from various data sources presents a systematic approach to overcome limitations arising from the subjective nature of pilot reports.

Data

PIREP

The IEM (Iowa Environmental Mesonet) database consists of PIREP (Pilot Reports) data, which contains information reported by pilots during their flights about atmospheric conditions they encounter. In this study, data obtained from the IEM database was utilized. For proper matching with meteorological data, information on latitude, longitude, altitude, and time of each turbulence observation is required. The IEM database provides all this information and also contains details about the types and severity of turbulence reported by pilots. In this study, the data was filtered according to specific criteria:

1. Only high-altitude data at pressure levels of 200, 225, 250, 300, and 350 hPA was selected.
2. Data covering the time period between 2022 and 2024 was used.
3. Geographically, the US airspace between 50°N and 24°N latitude, and -125°E and -67°E longitude was examined.
4. Among turbulence types, only Clear Air Turbulence (CAT) related data was selected.

This filtering allowed the study to focus specifically on Clear Air Turbulence cases occurring at high altitudes within US airspace.

Between 2022 and 2024, a total of 19,213 Clear Air Turbulence (CAT) cases were recorded. These CAT events were classified into five different categories based on their intensity: LGT, LGT-MOD, MOD, SEV-MOD, SEV.

[

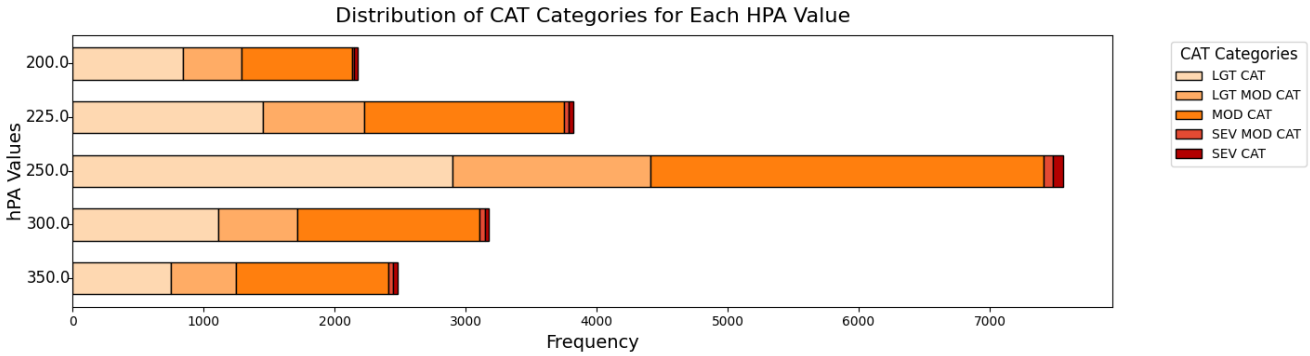


Fig. 1: Distribution of CAT Categories for Each HPA Value in US Airspace between 2022 and 2024.

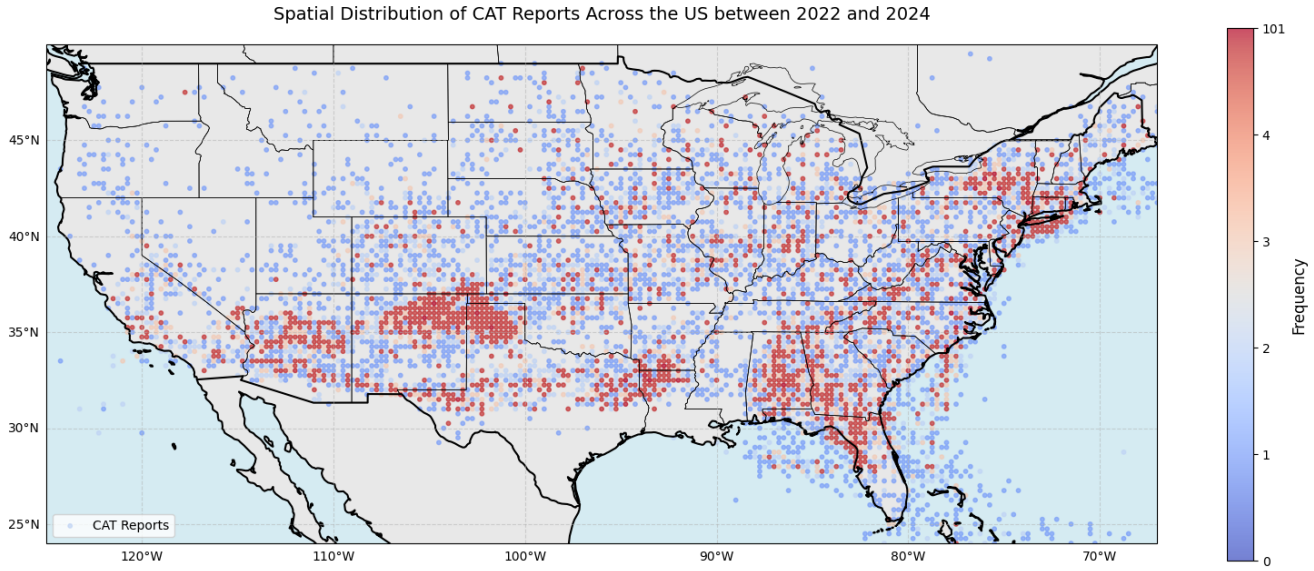


Fig. 2: Spatial Distribution of CAT Reports Across the US between 2022 and 2024.

2] Creating a balanced dataset is crucial for the effective application of machine learning models. Therefore, to balance the dataset, 19,213 NEG (negative - no turbulence) cases were also selected to match the 19,213 positive CAT cases, resulting in a balanced dataset consisting of a total of 38,426 cases.

Figure 1, shows the distribution of CAT events by severity categories recorded at different pressure levels between 2022 and 2024. The highest frequency of CAT events is observed at the 250 hPA level, where all severity categories are represented. A particularly notable point is that MOD CAT events are significantly higher at the 250 hPA level.

Figure 2, illustrates the spatial distribution of CAT reports across US airspace between 2022 and 2024. The visualization reveals distinct regional clustering of CAT events.Notably, high-frequency CAT reports are concentrated in the Southwest region (Arizona, New Mexico, and surrounding areas),the South (intersection of Texas, Oklahoma, and Arkansas), along the Eastern seaboard (from the Carolinas to Florida), and in the Northeast corridor (particularly around New York). This distribution likely correlates with the mountainous terrain features, jet stream pathways, and air traffic density in these regions.

Figure 3, illustrates the seasonal distribution of turbulence categories from 2022 to 2024. The graph clearly demonstrates a seasonal cycle in turbulence reports. Winter and spring months show a marked increase in MOD CAT reports, with the 2023-2024 winter season reaching the highest number of MOD CAT reports at approximately 1100 incidents. LGT CAT reports follow a similar seasonal pattern, tending to increase during winter and spring months. SEV CAT and SEV MOD CAT categories remain relatively low throughout the year but exhibit slight increases during winter seasons.

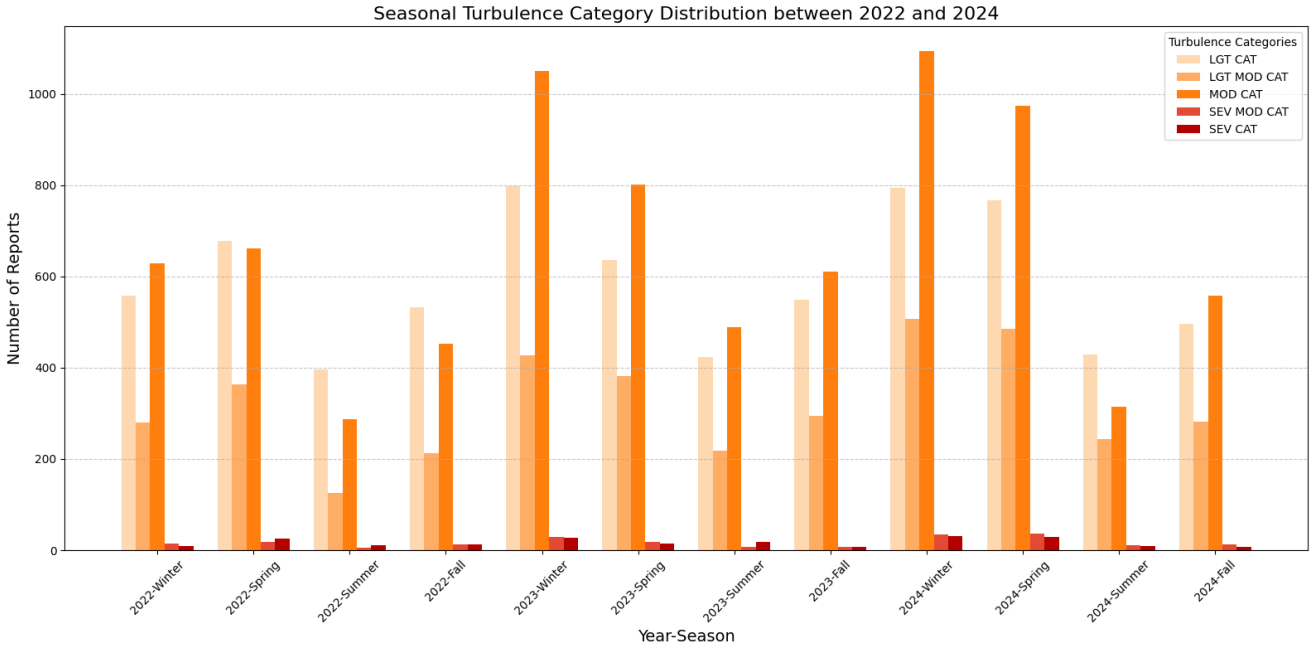


Fig. 3: Seasioanl Tubulence Category Distribution Across the US between 2022 and 2024.

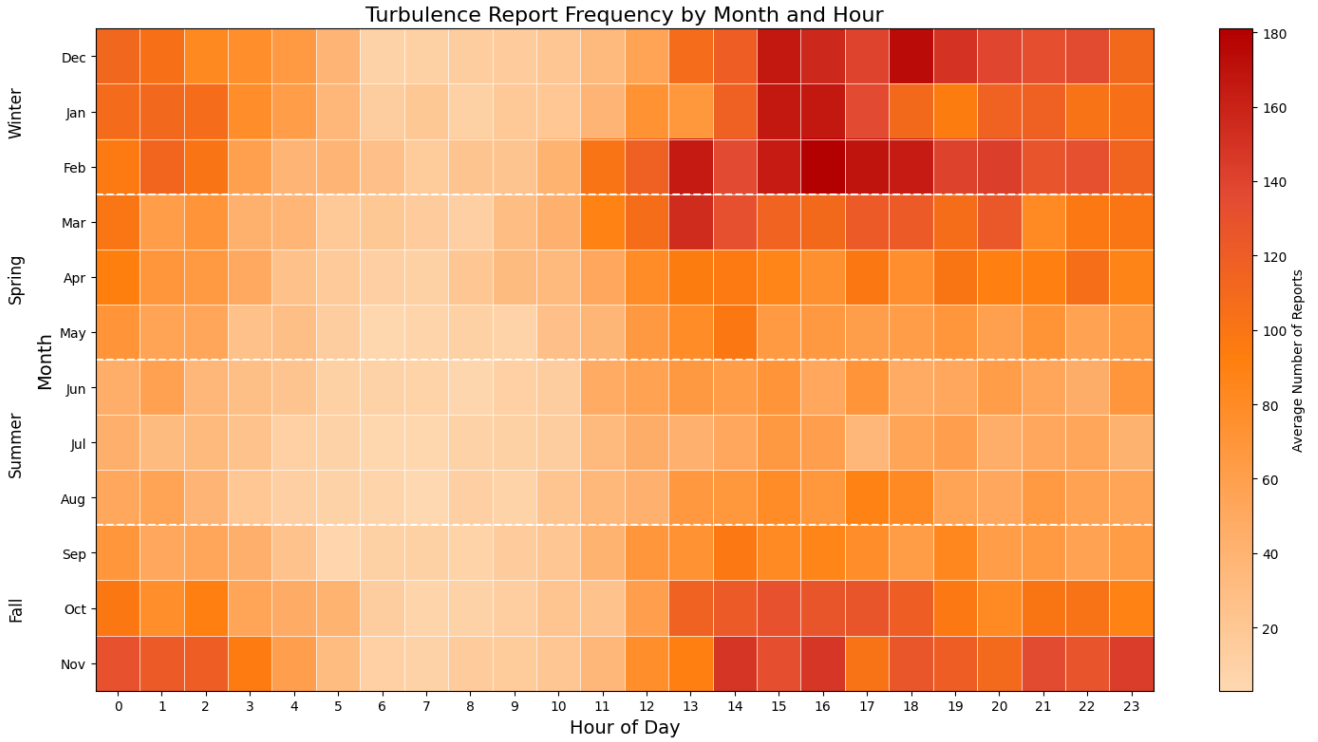


Fig. 4: Turbulence Report Frequency by Month and Hour

Summer months show a notable decrease in all turbulence categories, suggesting that summer conditions are less conducive to CAT formation. This seasonal pattern can be attributed to jet stream activity and stronger atmospheric frontal systems during winter months.

Figure 4, illustrates the frequency of turbulence reports throughout the year, broken down by month and hour of day. Winter months (December, January, February) show a significant increase in turbulence reports, particularly during evening hours (15:00-21:00). During these periods, the number of reports reaches 140-180.Across all months, early morning hours and late afternoon/evening hours show increased turbulence reports. The period between 14:00-20:00 is consistently the most intense. Summer months (June, July, August) and midday hours (6:00-11:00) represent the periods with the fewest turbulence reports. Spring (March-May) and fall (September-November) months show a transition between winter and summer patterns. Turbulence reports begin to increase in November and start decreasing from March onward.

Based on this data, flight planning recommendations could include exercising additional caution during evening hours in winter months, while summer months and midday hours could be considered as periods with lower expected turbulence.

ERA 5 Data

The ERA 5 pressure level dataset is frequently preferred in turbulence analysis studies (Shao et al. 2024, Sharman et al. 2016, de Mello 2024). The ERA5 reanalysis dataset has a spatial resolution of 0.25 degrees latitude by 0.25 degrees longitude and a temporal resolution of 1 hour. This research examines pressure levels of 200, 225, 250, 300, and 350 hPa, which correspond to the typical high-altitude flight levels used by aircraft. The study covers data from 2022 to 2024 and focuses on the region between 50°N and 24°N latitude, and -125°E and -67°E longitude.

The pressure levels database in ECMWF includes important variables related to turbulence, such as geopotential, relative humidity, temperature, potential vorticity, and wind components (U, W, V).

Due to the complex nature of turbulence, rather than attempting to achieve a complete representation of turbulence, there is a need for indicators that signify the presence of turbulence. The turbulence indicators used in this study are comprehensively summarized in Table 2. These indicators, based on various atmospheric parameters, allow us to evaluate the potential for turbulence.The selected parameters can be fundamentally categorized into two main groups: dynamic wind components and thermal effects.

Parameters representing thermal effects include components such as Theta (potantial temperature) and TR (temperature gradient/Richardson number). These parameters measure temperature changes in the atmosphere, modeling the impact of thermal instabilities on turbulence formation. Regions where temperature gradients change abruptly are particularly critical for clear air turbulence (CAT) formation.

Dynamic wind components include parameters such as wind shear, Dutton index, and DVT, which characterize both spatial and temporal variations of wind. These parameters measure the complex dynamic behaviors of air masses in the atmosphere, reflecting the energy potential necessary for turbulence formation.

Exploratory data analysis revealed that turbulence diagnostic parameters in the ERA5 dataset exhibit a non-Gaussian distribution. Figure 5 illustrates the distribution of turbulence diagnostics across severity levels

Table 2 Atmospheric turbulence diagnostic parameters and their formulations.

Diagnostic	Description	Unit	Algorithm	Reference
Ri	Negative Richardson number	—	$\mathrm{Ri} = -\frac{N^2}{\mathrm{VWS}^2}$	Oard (1974)
TI3[a]	Variant 3 of Elrod's turbulence index	s^{-1}	TI3 = TI1 + DVT	Lee et. al. (2022)
PVGRAD[b]	Potential vorticity gradient	PVU km^{-1}	$\mathrm{PV}_{\mathrm{grad}} = \sqrt{150\left(\nabla_h \mathrm{PV}\right)^2 + \left(\frac{\partial \mathrm{PV}}{\partial z}\right)^2}$	Homeyer et. al. (2021)
VORSQ	Relative vorticity squared	s^{-2}	$\mathrm{VORSQ} = \left\lvert \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right\rvert^2$	Sharman et al. (2006)
DVT[c]	Divergence Tendency	s^{-1}	$\mathrm{DVT} = \mathrm{C}\left\lvert \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)_{t2} - \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)_{t1} \right\rvert$	Lee et. al. (2022)
Dutton[d]	Dutton's empirical index	—	$\mathrm{Dutton} = 1.25\lvert\mathrm{HWS}\rvert + 0.25\mathrm{VWS}^2 + 10.5$	Dutton (1980)
|DIV|	Magnitude of horizontal divergence	s^{-1}	$\lvert\mathrm{DIV}\rvert = \left\lvert \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right\rvert$	Sharman et al. (2006)
TR	Horizontal temperature gradient/Richardson number	K m^{-1}	$\mathrm{TR} = \frac{\lvert \frac{\partial t}{\partial x} + \frac{\partial t}{\partial y} \rvert}{Ri}$	Kim et al. (2018)
θ	Potential temperature	K	$\theta = T\left(\frac{p_0}{p}\right)^{R/c_p}$	Holton et. al. (2012)
Wspeed	Wind speed	m s^{-1}	$\mathrm{Wspd} = (u^2 + v^2)^{1/2}$	Endlich (1964)
TI2[e]	Variant 2 of Elrod's turbulence index	s^{-2}	TI2 = VWS × (DEF-DIV)	Elrod and Knapp (1992)
WINDS/RI[f]	Wind Speed^2/Richardson Number	$\mathrm{m}^2\ \mathrm{s}^{-2}$	$\mathrm{WINDS/RI} = \mathrm{Wspeed}^2/\mathrm{Ri}$	Holton et. al. (2012)
VWS	Vertical shear of horizontal wind	s^{-1}	$\mathrm{VWS} = \left[\left(\frac{\partial u}{\partial z}\right)^2 + \left(\frac{\partial v}{\partial z}\right)^2\right]^{1/2}$	Endlich (1964)
N^2	Brunt-Väisälä frequency squared	s^{-2}	$\mathrm{N}^2 = \frac{g}{\theta}\frac{\partial \theta}{\partial z}$	Sharman et al. (2006)

In this table, several parameters require additional clarification. The variable TI1 refers to Elrod's original turbulence index, calculated as the product of vertical wind shear (VWS) and flow deformation (DEF). Potential vorticity (PV) is defined by the equation $-g(\frac{\partial \theta}{\partial p})(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} + f)$, where f represents the Coriolis parameter. For the divergence tendency (DVT) calculation, a constant scaling factor of $C = 0.01$ is applied. Horizontal wind shear (HWS) is expressed as $\frac{1}{V^2}(v\frac{\partial u}{\partial x} - u\frac{\partial u}{\partial y} + v\frac{\partial v}{\partial x} - u\frac{\partial v}{\partial y})$, while flow deformation (DEF) equals $\sqrt{\mathrm{DSH}^2 + \mathrm{DST}^2}$, combining stretching deformation ($\mathrm{DSH} = \frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}$) and shearing deformation ($\mathrm{DST} = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}$).

using box plots. For comparative visualization purposes only, we normalized the values in this figure, while maintaining non-normalized raw data throughout the actual study.

The most significant finding in the analysis results is the different distribution patterns of turbulence parameters across severity categories. Parameters such as $n2$,*wing loading ref*, and t exhibit high normalized values across all turbulence intensity levels, demonstrating their fundamental position in turbulence characterization regardless of severity level. In contrast, parameters including *richardson*, *vorsq*, and *ti*3 show notably low values, with the Richardson number's consistently low values confirming its known inverse relationship with atmospheric instability and turbulence potential. The increasing values of wind-related parameters such as *wind shear* from MOD CAT to MOD SEV CAT indicate their potential importance in distinguishing between severity levels, while revealing the complex, non-linear structure in turbulence manifestation.

The non-normalized raw data used in our study preserves the actual physical values of the parameters and their natural relationships with each other. This approach enables the creation of better representations of the physical foundations of turbulence formation mechanisms.

BADA

BADA (Base of Aircraft Data) is a comprehensive aircraft performance model developed and maintained by EUROCONTROL. This database contains detailed operational characteristics for various aircraft types, including performance parameters, aerodynamic coefficients, and engine thrust properties. As an alternative to classification methods that are independent of aircraft type, such as GTG, this study investigates the potential effects of incorporating aircraft type and aerodynamic features from pilot reports into the model.

Pilot reports (PIREPs) generally include aircraft type information, which can be matched with the relevant aerodynamic properties in the BADA database. This integration offers a more comprehensive prediction approach compared to traditional aircraft-type-independent classification methods such as GTG by incorporating how aircraft type and aerodynamic properties affect the turbulence experience into the model.

Aircraft-Specific Aerodynamic Parameters

This study analyzes the influence of aircraft-specific aerodynamic parameters on turbulence response. We utilize BADA (Base of Aircraft Data) parameters to calculate key aerodynamic metrics, establishing relationships between aircraft type and turbulence intensity perception.

Mathematical Formulation

The aircraft-specific parameters were calculated using the following equations:

$$L_{max} = \frac{1}{2}\rho V^2 S(C_{L,bo} \cdot B_c) \quad \text{(Maximum lift force)}$$

$$n_{ref} = \frac{L_{max}}{m_{ref} \cdot g} \quad \text{(Load factor)}$$

$$F_d = \frac{1}{2}\rho V^2 S C_d \quad \text{(Drag force)}$$

$$\frac{L}{D} = \frac{L_{max}}{F_d} \quad \text{(Lift-to-drag ratio)}$$

$$\frac{C_L}{C_d} = \frac{C_{L,bo} \cdot B_c}{C_d} \quad \text{(Lift-to-drag coefficient ratio)}$$

$$AR = \frac{b^2}{S} \quad \text{(Aspect ratio)}$$

$$\frac{W}{S} = \frac{m_{ref} \cdot g}{S} \quad \text{(Wing loading)}$$

Where:

- ρ: Air density [kg/m^3]
- V: True airspeed [m/s]
- S: Wing area [m^2]
- $C_{L,bo}$: Basic lift coefficient
- B_c: Buffet coefficient
- m_{ref}: Reference mass [kg]
- g: Gravitational acceleration [9.81 m/s^2]
- C_d: Drag coefficient
- b: Wingspan [m]

Methods

Turbulence Diagnostic Parameters and Data Preparation

For the input of our machine learning model, various turbulence diagnostic parameters with proven significance in aviation literature have been selected. Comprehensive descriptions of these parameters are presented in detail in Table 2. Data obtained from Pilot Reports (PIREPs) provides latitude, longitude, altitude, and time information, and turbulence diagnostic parameters in ERA5 are labeled with these. In the modeling phase, instances where Clear Air Turbulence (CAT) was present were labeled as "1" and instances where it was absent were labeled as "0" using a binary classification approach.

One of the distinctive aspects of our study is the inclusion of aircraft aerodynamic properties in the model, in addition to traditional turbulence diagnostic parameters. This integration was accomplished using data provided from EUROCONTROL's BADA (Base of Aircraft Data) database. The aircraft type information provided by PIREPs was matched with corresponding aerodynamic properties in the BADA database, enabling a more comprehensive turbulence prediction model.

Data Categorization and Analysis Scenarios

In our research, we categorized our dataset in various ways to more deeply understand the effect of aerodynamic properties on turbulence detection. The data was divided into two categories based on turbulence intensity: "light to ligt moderate" and "moderate to severe." Additionally, cases where aerodynamic properties were included and cases where they were excluded were analyzed separately.When all these categorizations and general performance analyses are combined, a total of six different scenarios were examined in detail in our study.

In addition to RandomForest and XGBoost algorithms, which are predominantly used for CAT diagnosis in the literature, we also implemented AdaBoost and LightGBM algorithms. Furthermore, Logistic Regression was examined as a potential alternative to decision tree algorithms.The entire dataset was randomly split into 70% training and 30% testing data for model training and evaluation, and all models were analyzed

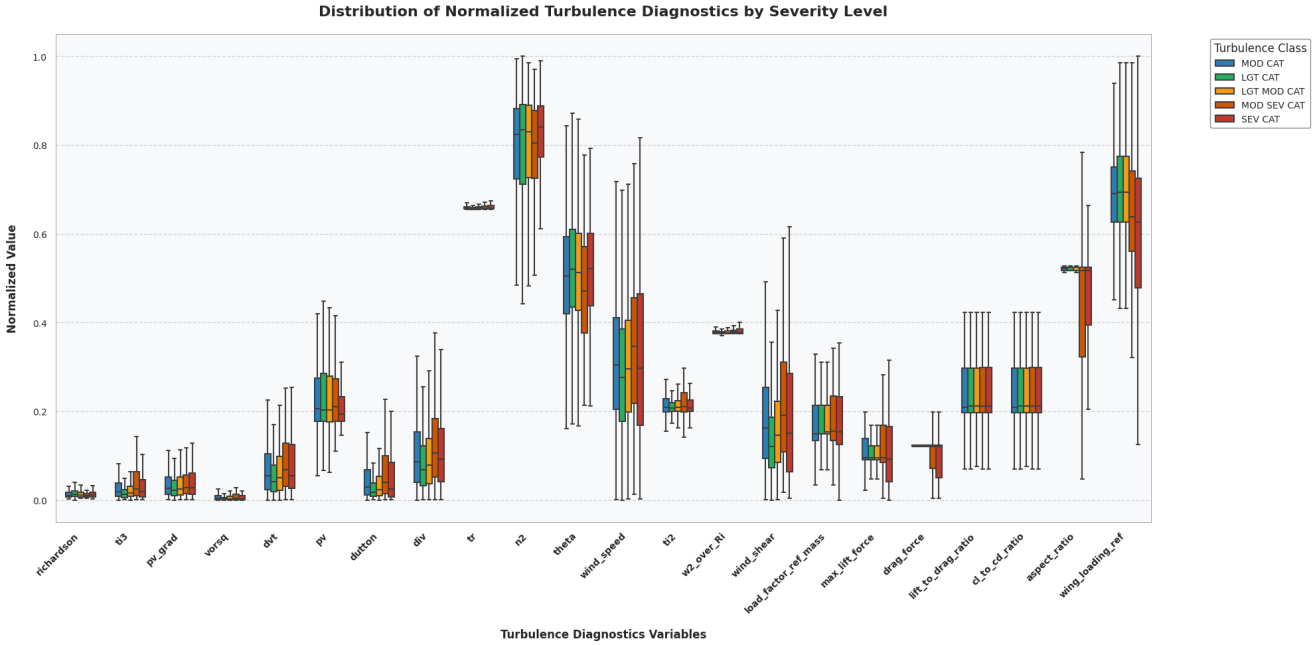


Fig. 5: Distribution of Normalized Turbulence Diagnostics by Severity Level

Table 3 Metrics used for evaluating model performance.

Metric	Formula	Description
POD (Probability of Detection)	$\frac{TP}{TP+FN}$	Rate of correctly detected turbulence events. Higher values indicate better detection of hazardous areas, enhancing flight safety.
FAR (False Alarm Ratio)	$\frac{FP}{TP+FP}$	Proportion of false turbulence predictions. Lower values reduce unnecessary route changes and improve operational efficiency.
CSI (Critical Success Index)	$\frac{TP}{TP+FP+FN}$	Overall success in predicting turbulence events, considering both false alarms and missed events. Balances safety and operational efficiency.
TSS (True Skill Statistic)	$\frac{TP}{TP+FN} - \frac{FP}{FP+TN}$	Balanced measure of prediction ability for both turbulent and non-turbulent conditions. Ranges from -1 to 1, with 1 being perfect prediction. Effective for imbalanced datasets.

using these data partitions.

Model Evaluation

For a single classifier model, the outputs can be categorized into four types: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). These outputs are used as calculation metrics for ROC (Receiver Operating Characteristic) curves and AUC (Area Under Curve) to evaluate model performance. This approach is frequently preferred in the literature for evaluating CAT prediction models (Cai et al. 2023; Hu et al. 2022; Kim et al. 2018; Lee et al. 2020; Sharman et al. 2006).

In addition, POD (Probability of Detection), FAR (False Alarm Ratio), CSI (Critical Success Index), and TSS (True Skill Statistic) metrics, which are commonly used in the literature for CAT prediction, were also employed in the evaluation (Sharman et al. 2016, Pearson et al. 2017, Mu˜noz- Esparza et al. 2020). The formulations for these metrics are presented in detail in Table 3.

POD indicates the rate at which turbulence events that actually occur are correctly detected by the model. A high POD value enhances flight safety by ensuring pilots are informed of potentially hazardous areas.

FAR represents the proportion of events predicted as turbulence by the model but which do not actually occur. A low FAR value improves operational efficiency by preventing unnecessary route changes and fuel consumption.

CSI comprehensively evaluates the model's success in correctly predicting turbulence events. It takes into account both false alarms and missed turbulence events, thus reflecting the balance between safety and efficiency in aviation operations.

TSS assesses the model's ability to correctly predict both turbulent and non-turbulent air conditions in a balanced manner. This metric is particularly important in imbalanced datasets like turbulence prediction, as turbulence events occur infrequently.

The AUC score of 0.5 represents the lower performance limit producing random results, while a score of 1 represents the ideal classifier. In this context, the performance of the models developed for high-altitude CAT diagnosis has been comprehensively evaluated using these metrics.

Results

In this study, the performance of XGBoost, LightGBM,

Table 4 Performance Comparison of Machine Learning Models Without and With Aerodynamic Features

Turbulence	Model	Without Aerodynamic Features					With Aerodynamic Features				
		POD	FAR	CSI	TSS	AUC	POD	FAR	CSI	TSS	AUC
All Categories	Random Forest	0.790	0.175	0.677	0.623	0.887	0.785	0.170	0.676	0.624	0.889
	XGBoost	0.796	**0.154**	**0.695**	**0.652**	**0.901**	0.809	**0.158**	**0.703**	**0.657**	**0.904**
	LightGBM	**0.800**	0.160	0.694	0.648	0.900	**0.815**	0.166	0.701	0.652	0.902
	AdaBoost	0.790	0.161	0.692	0.641	0.893	0.770	0.161	0.682	0.633	0.893
	Linear Regression	0.633	0.247	0.524	0.425	0.767	0.624	0.273	0.505	0.389	0.734
MOD-SEV	Random Forest	0.828	**0.141**	0.728	0.691	0.919	0.849	0.155	0.735	0.693	0.919
	XGBoost	0.845	0.139	0.743	0.708	**0.928**	**0.866**	0.150	**0.753**	**0.716**	**0.928**
	LightGBM	**0.878**	0.161	**0.751**	**0.709**	**0.928**	0.857	0.154	0.741	0.701	0.923
	AdaBoost	0.848	0.149	0.742	0.702	0.921	0.839	**0.140**	0.741	0.695	0.923
	Linear Regression	0.682	0.205	0.580	0.506	0.816	0.651	0.211	0.554	0.477	0.786
LGT-LGTMOD	Random Forest	0.790	0.204	0.656	0.587	0.870	0.758	**0.188**	0.645	0.582	0.872
	XGBoost	**0.802**	0.196	**0.670**	**0.606**	**0.881**	**0.803**	0.197	**0.671**	**0.606**	**0.885**
	LightGBM	0.788	**0.190**	0.665	0.603	0.880	0.790	0.189	0.667	**0.606**	0.879
	AdaBoost	0.779	0.200	0.651	0.579	0.868	0.789	0.203	0.660	0.594	0.871
	Linear Regression	0.598	0.261	0.494	0.386	0.737	0.569	0.267	0.471	0.361	0.715

Table 5 Algorithm performance metrics across different turbulence intensity categories

Turbulence Category	POD	FAR	CSI	TSS	AUC
Neg					
σ_{XGBoost}	0.001	0.001	0.003	0.003	0.001
LGT					
σ_{XGBoost}	0.005	0.002	0.001	0.004	0.002
LGT-MOD					
σ_{XGBoost}	0.004	0.003	0.001	0.003	0.011
MOD					
σ_{XGBoost}	0.000	0.003	0.001	0.001	0.002
SEV-MOD					
σ_{XGBoost}	0.013	0.020	0.000	0.023	0.002
SEV					
σ_{XGBoost}	0.011	0.020	0.001	0.022	0.003

Note: To evaluate model stability and generalization capability, we implemented a k-fold cross-validation analysis for the baseline XGBoost model. Specifically, a fivefold approach was utilized (k = 5), maintaining consistency with our primary data partitioning strategy of 70% for training and 30% for testing datasets. This methodology allowed us to systematically assess the variability of statistical performance metrics across different data subsets, providing a robust evaluation of the model's predictive reliability.

Random Forest, AdaBoost, and Logistic Regression models for high-altitude clear air turbulence (CAT) prediction was comprehensively examined, taking into account the integration of aerodynamic features. The findings presented in Table 4 clearly demonstrate that gradient boosting-based models (XGBoost and LightGBM) exhibited the highest performance across all categories. XGBoost, which had an AUC value of 0.901 before the addition of aerodynamic features, maintained its leading position with an AUC value of 0.904 after the integration of these features. Particularly in the moderate to severe turbulence (MOD-SEV) category, the increase in the model's POD (Probability of Detection) value from 0.845 to 0.866 provided a significant improvement in the detection of dangerous turbulence events. This improvement contributes to more accurate detection of turbulence events by reducing false negative rates, presenting an important advancement toward enhancing operational safety.

The LightGBM algorithm also demonstrated similarly impressive performance. With the addition of aerodynamic features, the POD value in "All Categories" increased from 0.800 to 0.815, and the TSS (True Skill Statistic) rose from 0.648 to 0.652, exhibiting a balanced prediction capability. These results once again confirm the superiority of gradient boosting methods in modeling the non-linear dynamics of turbulence.

While the integration of aerodynamic features did not provide a consistent reduction in FAR (False Alarm Ratio) values overall, supportive effects on operational efficiency were observed in some models. For example, the AdaBoost algorithm successfully

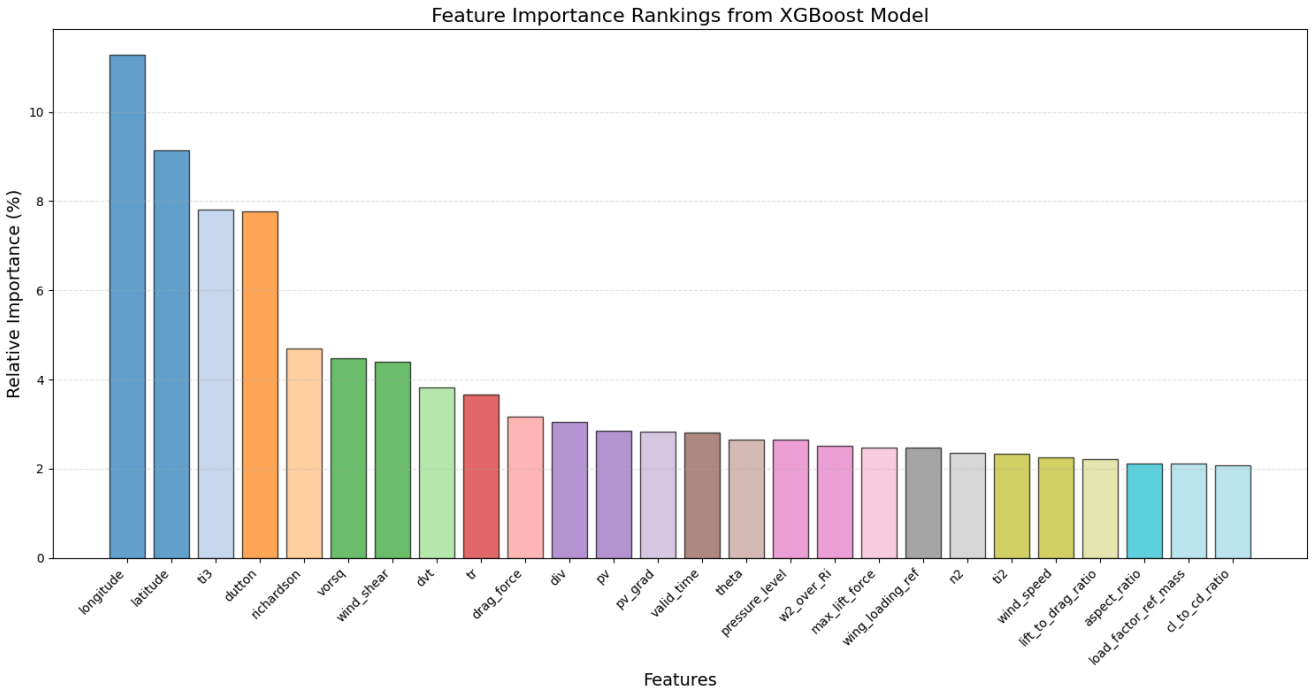


Fig. 6: All Categories XGBoost Relative Feature Importance

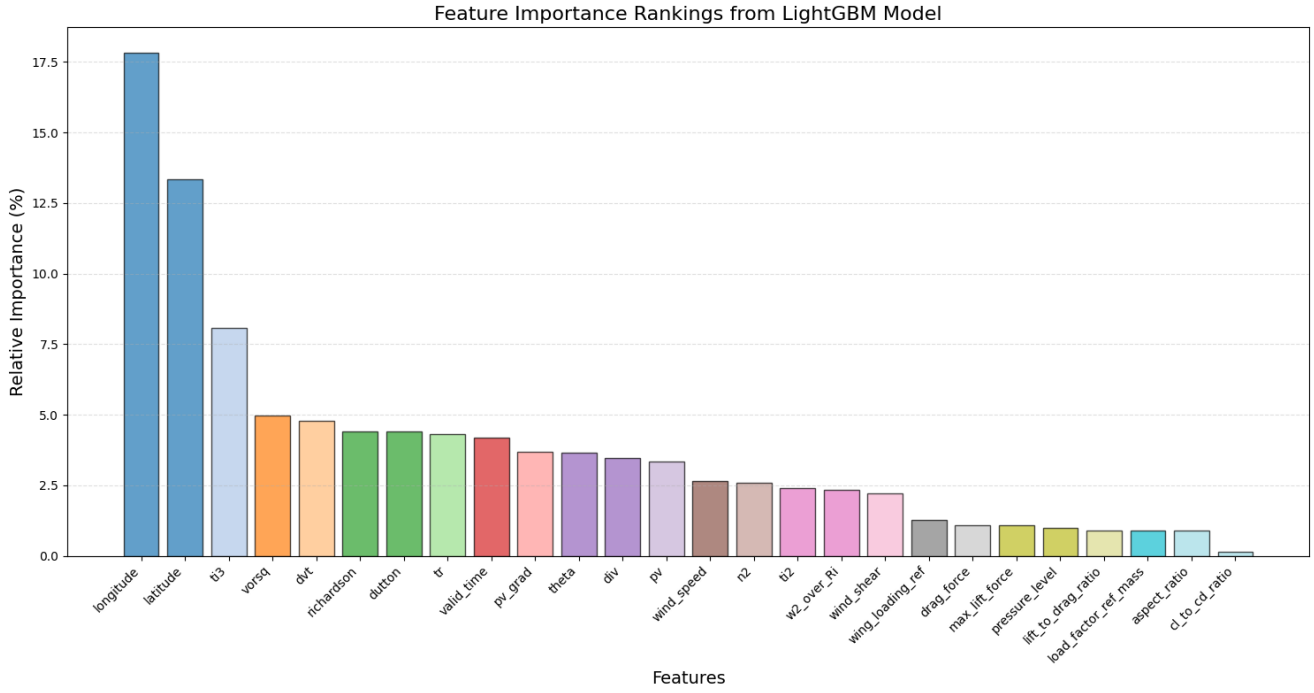


Fig. 7: All Categories LightGBM Relative Feature Importance

reduced FAR from 0.149 to 0.140 in the "MOD-SEV" category,demonstrating success in controlling false alarms. Similarly, the Random Forest model recorded a marginal improvement by reducing FAR from 0.175 to 0.170 in "All Categories." However, the increase of FAR from 0.154 to 0.158 in XGBoost shows that these features may affect the model's specificity in some cases.

The performance of Logistic Regression remained significantly lower compared to other models. The decrease of the AUC value from 0.767 to 0.734 in "All Categories" clearly demonstrates that the complex physical structure of turbulence cannot be captured with linear models. This situation once again highlights the inevitable superiority of ensemble methods in atmospheric events dominated by non-linear relationships.

The standard deviation values presented in Table 5 demonstrate that the performance metrics of the XGBoost model are quite consistent during the cross-validation process (σ_{AUC}: 0.001–0.003). According to the results in Table 4, the 0.021 increase in POD and 0.003 improvement in AUC observed in the MOD-SEV category with the addition of aerodynamic features are significantly larger than the corresponding standard deviations (σ_{POD}: 0.013; σ_{AUC}: 0.002). This supports that these improvements are statistically significant. Changes in FAR values are similar in magnitude to the standard deviations (σ_{FAR}: 0.020), suggesting that fluctuations in this parameter may be due to random variation. In conclusion, the significance of the increases in the model's key performance indicators emphasizes the effectiveness of aerodynamic feature integration.

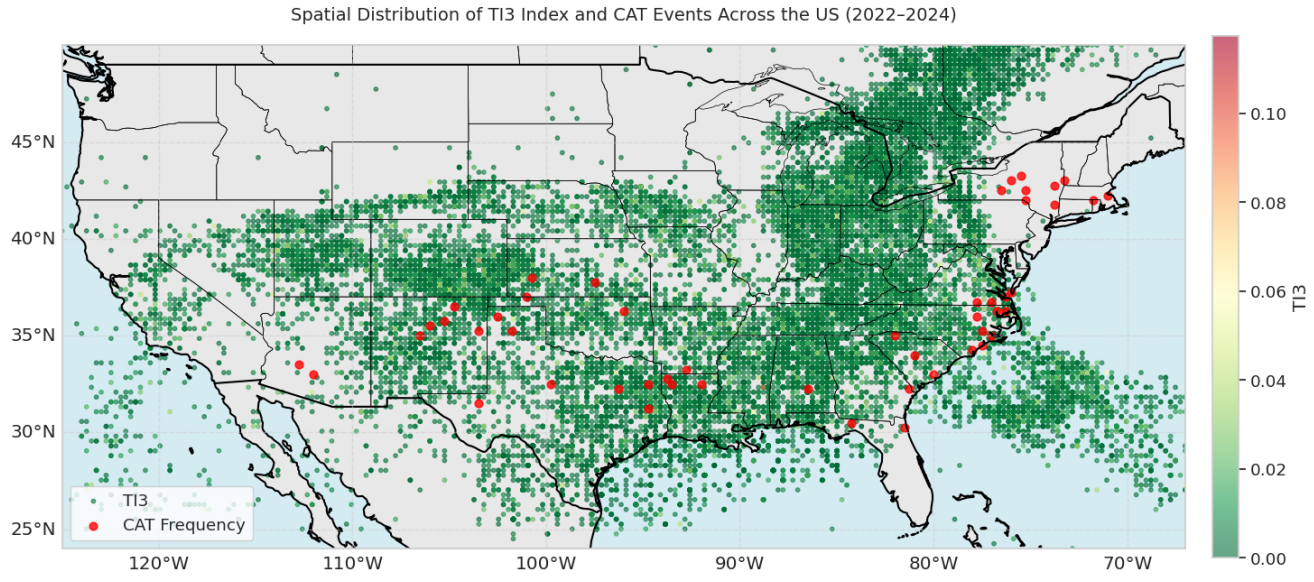


Fig. 8: Spatial Distribution of Ti3 Index and CAT Events Across the US between 2022 and 2024

Figures 6 and 7 show the feature importance rankings of LightGBM and XGBoost models, respectively. In both models, geographic coordinates (longitude and latitude) stand out as the most determinant features. Longitude has a relative importance of 17.5% in LightGBM and 11% in XGBoost. This demonstrates that turbulence events have a strong spatial relationship with topographic structures and regional meteorological dynamics.

While the atmospheric stability indicator *ti3* parameter ranks third in both models (LightGBM: 8.0%; XGBoost: 7.8%), the *dutton* index related to vertical wind shear is notable in XGBoost with an importance level of 7.8%. These findings indicate that atmospheric dynamic factors play a fundamental role in turbulence prediction.

Interestingly, while aerodynamic parameters are present in both models, they generally show relatively modest importance scores below 2.5%. Among these, *drag_force* ranks highest (3.2% in XGBoost), followed by *max_lift_force* and *wing_loading_ref*. This suggests that while aerodynamic features contribute to the predictive capabilities of the models, they provide supplementary rather than primary information for turbulence prediction. The consistency between LightGBM and XGBoost importance rankings reinforces the reliability of these findings; however, XGBoost shows a more balanced distribution of importance across features. This suggests that, as shown in Table 4, the XGBoost algorithm can extract some information from aerodynamic features that affect pilots' perception of turbulence.

Figure 8 presents the spatial relationship between the Ti3 index and Clear Air Turbulence events based on pilot reports in US airspace during the 2022-2024 period. Red dots represent regions where CAT reports are concentrated, while green dots represent areas where the Ti3 index is high. The distribution of these two parameters shows a distinct spatial overlap, except for around the New York area; this supports the physical connection between atmospheric dynamics and CAT.

Conclusion and Discussion

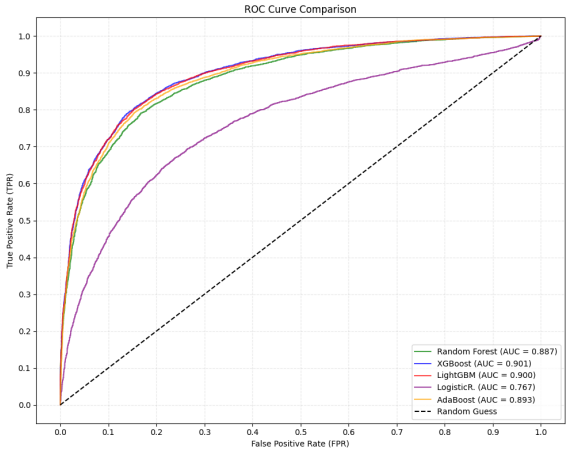
This study evaluates the performance of machine learning models in predicting high-altitude clear air turbulence (CAT) within the 200–350 hPa pressure levels and U.S. airspace. The AUC value of 0.904 achieved by the gradient boosting algorithm XGBoost is consistent with studies employing similar methodologies (e.g., Muñoz-Esparza et al., 2020: AUC 0.906), supporting its potential integration into operational systems. The dominant role of geographic coordinates (17.5%) and the Ti3 index clearly reflects CAT's relationship with topography and upper-tropospheric dynamics. This finding aligns with prior research emphasizing the criticality of regional meteorological factors in CAT formation (Muñoz-Esparza et al., 2020).

While the relatively limited contribution of aerodynamic parameters (3.2%) underscores the continued practicality of traditional aircraft-agnostic approaches, the increase in POD to 0.866 for MOD-SEV categories suggests their complementary role in risk management. The observed seasonal rise in CAT incidents during winter months, linked to jet stream activity and atmospheric instability, complements Shao et al. (2024)'s focus on convective systems.

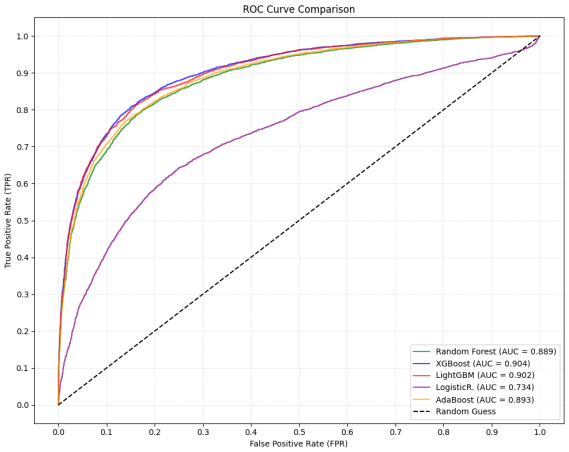
Key limitations include the dataset's restricted geographic scope and lack of aircraft type diversity. Future research could improve model generalizability through global-scale data integration and real-time telemetry. Additionally, innovative data balancing techniques to augment severe turbulence samples are recommended.

This study demonstrates the efficacy of machine learning models in high-altitude CAT prediction, highlighting the superiority of gradient boosting algorithms (XGBoost and LightGBM) in capturing non-linear atmospheric relationships. The limited yet strategic contribution of aerodynamic parameters offers quantitative advantages for early detection of severe turbulence in operational systems. The primacy of geographic and dynamic atmospheric variables emphasizes the need for regionally integrated prediction frameworks. By providing a data-driven approach to manage climate change-induced CAT increases, this research indicates pathways for further refinement through global data integration and algorithmic advancements.

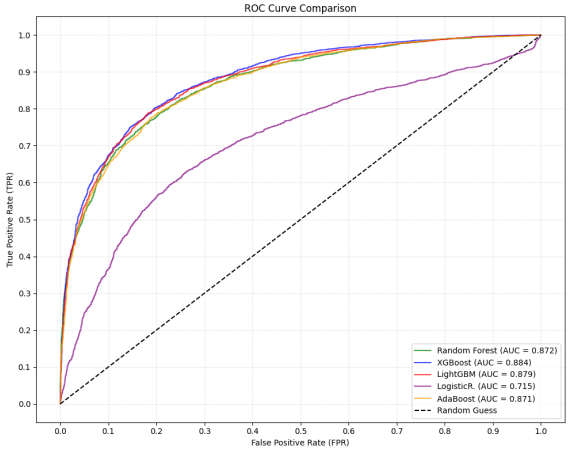
Acknowledgments We would like to thank the Presidency of Defence Industries of the Republic of Turkey and the SAYZEK-ATP program for the support and guidance provided within the scope of this study. The EUROCONTROL BADA dataset used in the study was provided by our technical mentor Tolga Başẗurk as part of the SAYZEK-ATP program.



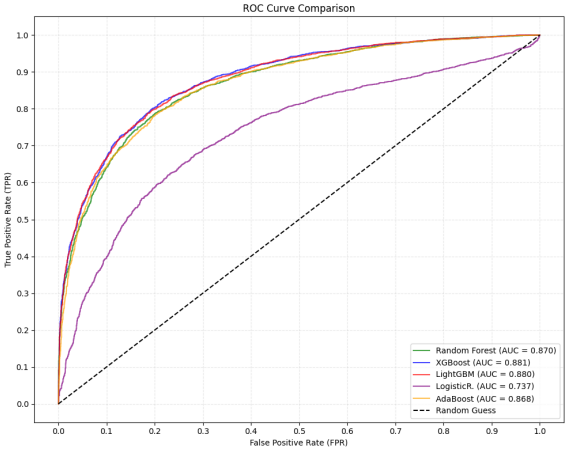
(a) All categories without aerodynamic features



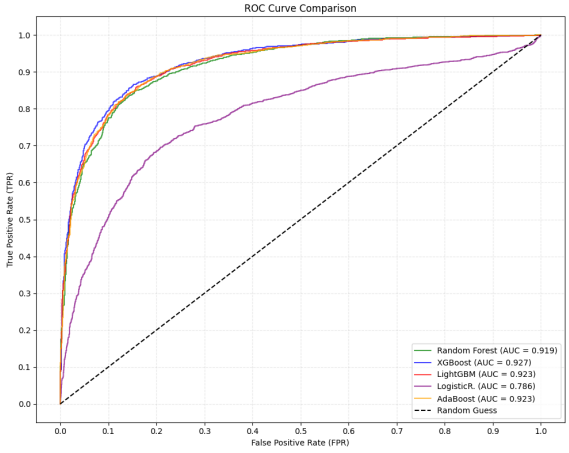
(b) All categories with aerodynamic features



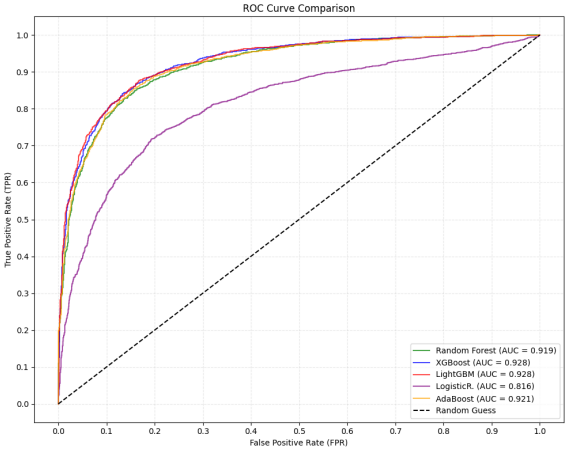
(c) Moderate or greater categories without aerodynamic features



(d) Moderate or greater categories with aerodynamic features



(e) Light and Light Mod categories without aerodynamic features



(f) Light and Light Mod categories with aerodynamic features

Fig. 9: "Comparison of ROC curves on different datasets. Each graph shows the classification performance of different machine learning models (Random Forest, XGBoost, LightGBM, Logistic Regression, and AdaBoost) along with their AUC values.