International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.37, September 2025

Machine Learning for Turbulence Risk Identification in U.S. Airspace using Open Flight and Weather Data

Godha Naravara
Department of EECS, Ohio University
USA

Ahmad Waseem Ghauri
Department of EECS, Ohio University
USA

Chad Mourning
Department of EECS, Ohio University
USA

ABSTRACT

Turbulence is a persistent and often unpredictable hazard in aviation, frequently occurring without visual or radar cues. This study presents a machine learning framework for identifying severe and extreme turbulence risk across U.S. airspace using only publicly available flight and weather data. The framework combines over 550,000 pilot reports with ERA5 reanalysis data to construct a large labeled dataset. It integrates anomaly-aware downsampling, synthetic oversampling, dimensionality reduction, and both unsupervised (K-Means) and supervised (XGBoost) modeling. In 10-fold cross-validation, the model achieved strong performance (recall = 0.91, F1 = 0.88) in detecting high-risk events. A real-world case study from February 2025 further illustrates the system's predictive capability. This work demonstrates the feasibility of operational turbulence identification using open-source data and interpretable learning techniques.

Keywords

Turbulence Identification, Machine Learning, Aviation Safety, ERA5 Reanalysis, PIREPs, XGBoost

1. INTRODUCTION

Turbulence is one of the leading causes of in-flight injuries and discomfort in commercial aviation. According to the National Transportation Safety Board (NTSB), turbulence accounted for over 70% of all serious nonfatal injuries in U.S. airline operations between 2009 and 2018, often impacting flight attendants during cruise phases without warning [14]. Turbulence can arise unexpectedly, especially in clear-air conditions where radar detection is limited, posing risks to passengers, crew, and aircraft integrity [7, 17]. Despite ongoing improvements in Numerical Weather Prediction (NWP) models and operational turbulence indices, such as those based on eddy dissipation rate (EDR) or wind shear, these systems continue to face limitations in detecting localized or transient atmospheric events [15, 13]. In particular, clear-air turbulence (CAT), which occurs without visible weather phenomena, remains difficult to forecast [16, 17]. In recent years, the increased availability of open-source atmospheric and flight data has opened new avenues for machine learning (ML) applications in aviation safety. These data-driven models complement forecasting tools by learning complex patterns from historical observations, especially for rare but high-impact events like severe turbulence [6]. This study builds on prior research, particularly the work by Mizuno et al. [12], who applied a combination of Principal Component Analysis (PCA), K-Means clustering, and Support Vector Machines (SVM) to detect turbulence risks near Matsumoto Airport in Japan using proprietary Quick Access Recorder (QAR) data.

This research relies on Pilot Reports (PIREPs) from the Iowa Environmental Mesonet[1] and ERA5 atmospheric reanalysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF)[2].

This paper makes the following key contributions:

—A large-scale, labeled dataset is curated by matching over 550,000 PIREPs with corresponding ERA5 weather data[2].

—A multi-stage ML pipeline is developed to address class imbalance, dimensionality reduction, and risk pattern discovery using anomaly filtering [11], synthetic oversampling [3], PCA [1] and K-Means clustering [8].

—Multiple supervised models are evaluated and compared, with XGBoost [4] identified as the most effective in predicting high-risk turbulence [10].

—A real-world case study on a high-severity day in 2025 is presented, demonstrating the practical utility of the proposed system.

Together, these contributions advance the use of open data and machine learning for operational turbulence forecasting.

2. RELATED WORK

Forecasting atmospheric turbulence remains a central challenge in aviation safety due to its high variability, nonlinear causes, and often limited observability. Traditional forecasting approaches rely on NWP models that estimate turbulence using features such as wind shear, eddy dissipation rate (EDR), and temperature gradients [15, 13]. While these methods are useful for large-scale trends, they often fail to capture localized or transient turbulence events, particularly in the case of CAT [17, 16].

[1] https://mesonet.agron.iastate.edu/request/gis/pireps.php
[2] https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels

As a result, there has been growing interest in data-driven models that can uncover hidden atmospheric patterns correlated with turbulence risk. However, much of the early machine learning (ML) work in this area has relied on proprietary datasets. For example, Mizuno et al. [12] proposed a turbulence-risk prediction pipeline combining PCA, K-Means clustering, and SVM, trained on QAR sensor data from commercial aircraft. Their unsupervised-to-supervised structure improved classification accuracy by identifying latent risk clusters in the meteorological feature space. However, the lack of public access to QAR data makes this approach difficult to replicate or scale beyond regional studies.

To overcome such data limitations, several researchers have turned to open-source alternatives. PIREPs are a widely available source of subjective, but in-situ, observations made by pilots during flight. Although reporting varies by aircraft type, turbulence sensitivity, and pilot judgment, studies have shown that PIREPs remain a viable proxy for turbulence events when carefully preprocessed [14]. Moreover, ERA5, a high-resolution reanalysis dataset, has become a standard for retrieving consistent atmospheric data across multiple variables and pressure levels [9].

Using these open datasets, de Mello et al. [6] applied Random Forest and Gradient Boosting models to predict clear-air turbulence using ERA5 and GFS features. Their results emphasized the importance of class imbalance handling and seasonal variability in improving model generalization. Similarly, Khattak et al. [10] used XGBoost in combination with SMOTE to detect wind shear events based on PIREPs at Hong Kong International Airport, demonstrating that ML pipelines can generalize to a range of flight hazards beyond turbulence.

Other studies have begun incorporating spatiotemporal learning techniques into operational aviation tools. For example, Chrit and Majdi [5] proposed a real-time turbulence nowcasting model for advanced air mobility that combines neural networks with regional weather features, demonstrating the importance of temporal dynamics in turbulence forecasting. While the current model operates on snapshot-level inputs, future extensions could integrate similar time-aware methods to enhance predictive accuracy, especially for rapidly evolving conditions.

Across these studies, several themes emerge. First, unsupervised methods such as PCA and clustering can help identify underlying structure in noisy atmospheric data [12]. Second, addressing class imbalance is critical for detecting rare but dangerous turbulence cases [3]. Third, most studies are limited in spatial or temporal scope and lack integration of both pilot observations and reanalysis data at scale.

The present study synthesizes these threads by building a turbulence identification pipeline that uses only open-access data sources, with PIREPs for labeling and ERA5 for weather features across the full U.S. airspace. The framework retains the unsupervised-to-supervised structure of Mizuno et al. but replaces proprietary QAR data with scalable alternatives. A class-balancing strategy is introduced that combines Isolation Forest–based anomaly-aware downsampling [11] with SMOTE [3], and incorporates risk clustering into supervised learning. Finally, the model is evaluated on both cross-validation metrics and a real-world high-risk case, illustrating its applicability for operational forecasting.

3. DATA AND PREPROCESSING

This study integrates two open-access datasets, PIREPs and ERA5 reanalysis data, to construct a large-scale, labeled dataset for turbulence-risk identification in U.S. airspace.

3.1 PIREPs: Turbulence Labeling

PIREPs are short in-flight observations voluntarily submitted by pilots, often describing turbulence, icing, or weather anomalies. Over 1.1 million PIREPs for the year 2024 were retrieved from the Iowa Environmental Mesonet (IEM) in shapefile format. Each report includes spatial coordinates, timestamp, altitude (reported as flight level), and textual annotations describing the observed turbulence. Figure 1 presents a geographic distribution of the PIREP reports.

To construct usable turbulence labels, the following preprocessing steps were applied:

—Extracted only records containing valid turbulence descriptors and altitude information.

—Converted flight levels to pressure altitude using MetPy's standard atmosphere function[3], which internally applies the barometric formula under ICAO Standard Atmosphere assumptions.

—Standardized turbulence descriptions into five categories: `NEG` (no turbulence), `LGT` (light), `MOD` (moderate), `SEV` (severe), and `EXTRM` (extreme).

—Removed duplicate or malformed entries.

This process yielded approximately 550,000 labeled turbulence reports across the continental U.S., Alaska, and Oceanic regions. To account for seasonal variability, each record was tagged with a `Season` feature derived from its timestamp (Winter, Spring, Summer, Fall).

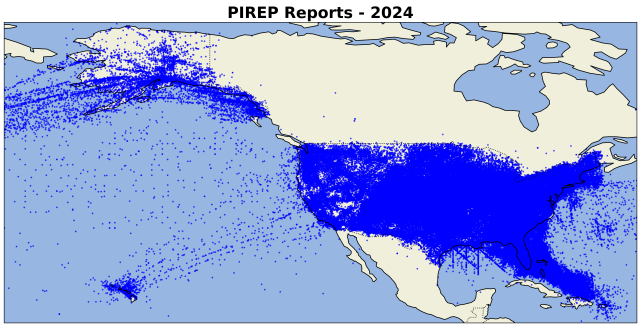

Fig. 1: Geographic distribution of over 1.1 million raw PIREPs collected during 2024. Each point corresponds to a pilot-submitted report within U.S. airspace, highlighting nationwide coverage.

3.2 ERA5: Meteorological Features

For each turbulence report, corresponding atmospheric conditions were retrieved from ERA5, a global reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) [9]. ERA5 provides hourly gridded estimates of

[3] `https://unidata.github.io/MetPy/latest/api/generated/metpy.calc.height_to_pressure_std.html`

meteorological variables across 37 pressure levels at 0.25° geodetic spatial resolution.

Monthly GRIB files for 2024 were queried using the Copernicus Climate Data Store API, and each PIREP was matched to the nearest timestamp, location, and pressure level. To minimize spatial interpolation errors, only reports falling within ERA5's defined geographic bounds over U.S. airspace were retained.

The selected ERA5 variables were chosen based on prior studies that linked them to turbulence generation mechanisms [12, 6]. The full list of weather features used in modeling is shown in Table 1, including wind components (u, v), vertical velocity, temperature, humidity, geopotential height, cloud cover, and potential vorticity. Derived features such as wind speed and wind direction were computed using vector transformations. All weather variables were standardized using z-score normalization prior to modeling.

3.3 Labeling Strategy and Class Distribution

To frame the task as binary classification, only reports labeled as `NEG` (non-turbulent) or belonging to the high-risk classes `SEV` and `EXTRM` were selected. Light and moderate turbulence reports were excluded due to their subjectivity and inconsistency across pilot experience and aircraft type. The final class distribution was:

—Class 0 (non-turbulent): ∼27,100 samples

—Class 1 (turbulent): ∼12,800 samples

To reduce geographic and temporal bias in training, non-turbulent reports were filtered to ensure they occurred within a ±1 hour window, within ±2 flight levels (approximately 2000 feet), and within a 2° latitude/longitude radius of nearby severe or extreme events. This constraint ensured that Class 0 samples reflected similar meteorological conditions and operational regions as Class 1 cases, minimizing confounding due to time, location, or aircraft routing. Each row in the final dataset includes normalized ERA5 weather features, categorical metadata (season, aircraft type, ARTCC region), and a binary turbulence label.

3.4 Feature Engineering and Metadata Encoding

Additional categorical features were incorporated to capture aircraft sensitivity and regional weather patterns:

—`Aircraft Type`: Numerically encoded from 3,225 unique aircraft identifiers that reported turbulence.

—`ARTCC Region`: Encoded from 21 Air Route Traffic Control Centers managing U.S. sectors.

The complete list of meteorological and metadata features used for training and evaluation is shown in Table 1. The final feature set included 24 meteorological variables, 2 encoded metadata fields, and 1 binary target label. These records served as inputs to subsequent preprocessing and modeling steps, described in Section 4.

4. METHODOLOGY

A modular machine learning pipeline was designed to process open aviation and weather data and generate interpretable turbulence-risk identifications. The pipeline consists of five stages: class imbalance handling, dimensionality reduction, unsupervised risk clustering, model training with cross-validation, and final evaluation.

Table 1. : Summary of final features used for modeling turbulence risk.

Feature Name	Units / Encoding	Description
Meteorological Features (from ERA5)		
Wind Speed	m/s	Magnitude of horizontal wind vector
Wind Direction	m/s	Wind flow direction from u and v components
U-component of Wind	m/s	East–west wind velocity
V-component of Wind	m/s	North–south wind velocity
Vertical Velocity (Omega)	Pa/s	Updraft or downdraft air motion
Vertical Wind Shear	(m/s)/km	Rate of wind change with altitude
Temperature	K	Air temperature at pressure level
Relative Humidity	%	Moisture content relative to saturation
Specific Humidity	g/kg	Absolute moisture content
Geopotential Height	m^2/s^2	Height of pressure surface
Potential Vorticity	$K{\cdot}m^2/kg{\cdot}s$	Atmospheric stability indicator
Relative Velocity	1/s	Local air rotation
Divergence	1/s	Horizontal airflow divergence
Cloud Fraction	Unitless	Proportion of cloud coverage (0–1)
Specific Cloud Liquid Water	kg/kg	Liquid water content in clouds
Specific Cloud Ice Water	kg/kg	Ice content in clouds
Metadata and Engineered Features		
Season	1–4 (int)	Winter (1), Spring (2), Summer (3), Fall (4)
Aircraft Type	0–3225 (int)	Encoded aircraft identifier
ARTCC Region	0–20 (int)	Encoded FAA regional airspace zone
Latitude, Longitude	degrees	Geographic coordinates
Altitude (Pressure Level)	hPa	Atmospheric pressure at report altitude
Risk Cluster Label	0/1	High-risk group from K-Means clustering
Binary Target	0/1	Non-turbulent (0), Severe/Extreme (1)

Feature values include ERA5 reanalysis weather variables and metadata extracted from PIREPs. These inputs were used in both unsupervised clustering and supervised turbulence classification tasks.

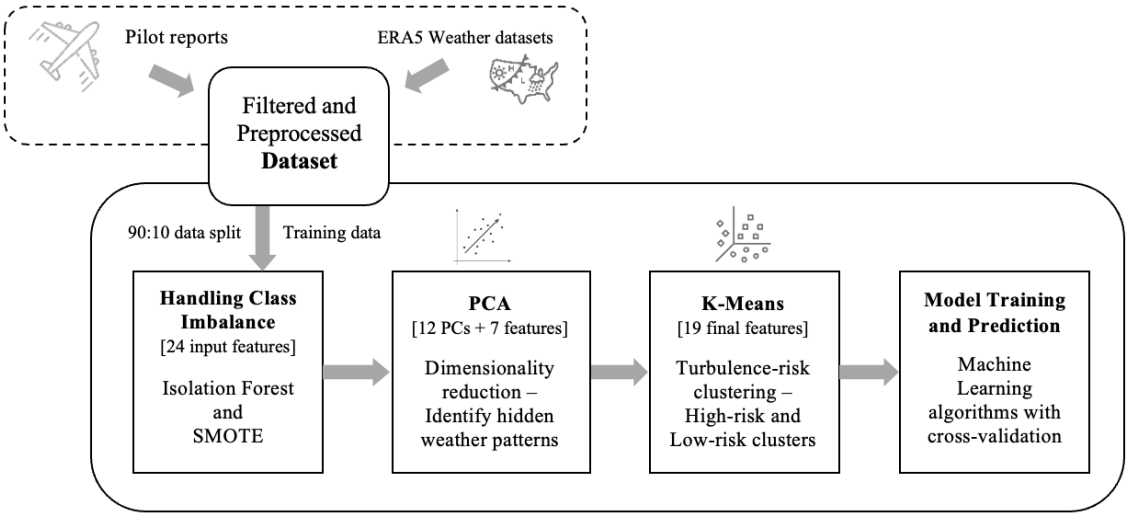

Fig. 2: Overview of the machine learning pipeline for turbulence-risk identification, including class balancing, dimensionality reduction, unsupervised clustering, and supervised classification.

4.1 Overview

Figure 2 provides a high-level overview of the end-to-end training pipeline. The input consists of labeled turbulence reports and ERA5 weather features. After preprocessing, the data passes through a series of transformations and modeling steps designed to address challenges such as class imbalance and feature redundancy.

4.2 Class Imbalance Handling

The dataset exhibited a natural skew, with non-turbulent reports (Class 0) outnumbering severe or extreme turbulence cases (Class 1) by a ratio of more than 2:1. To prevent model bias toward the majority class, a two-step class balancing strategy was applied:

—**Downsampling**: Redundant or low-diversity negative samples were filtered using Isolation Forest anomaly scores [11]. A blended sampling strategy retained diverse yet relevant negative examples by favoring mid-range anomaly scores over extreme outliers or typical samples.

—**Oversampling**: The SMOTE algorithm [3] was used to synthetically generate minority class examples by interpolating feature vectors from nearby severe/extreme samples. The minority class size was increased to 80% of the majority class, improving the model's exposure to high-risk conditions.

As shown in Figure 3, the anomaly score distribution guided downsampling to preserve diverse negative samples.

4.3 Dimensionality Reduction via PCA

To reduce feature redundancy and enable unsupervised risk discovery, PCA was applied to the 17 normalized weather features and altitude. The number of principal components was selected to retain 95% of the variance. These components replaced the original features for subsequent clustering and classification.

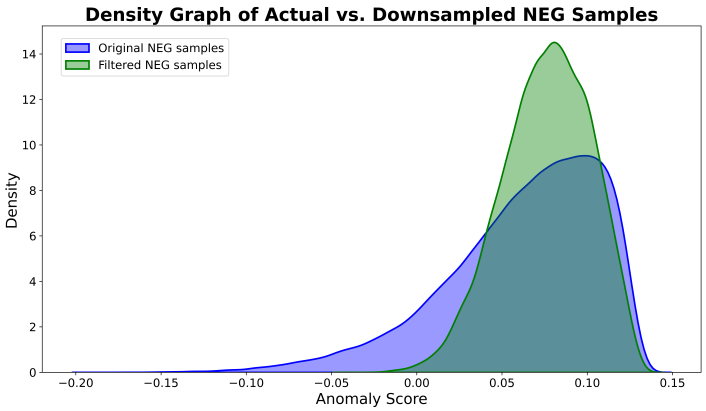

Fig. 3: Density plot of anomaly scores from Isolation Forest applied to non-turbulent samples. Mid-range scores were favored during downsampling to preserve diverse and informative negative cases.

4.4 Unsupervised Risk Clustering

Following Mizuno et al. [12], K-Means clustering was applied to the PCA-transformed data to identify latent structures in the feature space. The optimal number of clusters ($k = 3$) was selected using the Elbow Method [2]. One cluster showed a disproportionately high concentration of Class 1 reports and was labeled as high-risk; the others were labeled as low-risk. This cluster assignment was appended as a new feature to enhance model interpretability and support spatiotemporal analysis.

Figure 4 shows a 3D visualization of K-Means clusters in PCA-reduced space, highlighting a cluster enriched with severe turbulence reports.

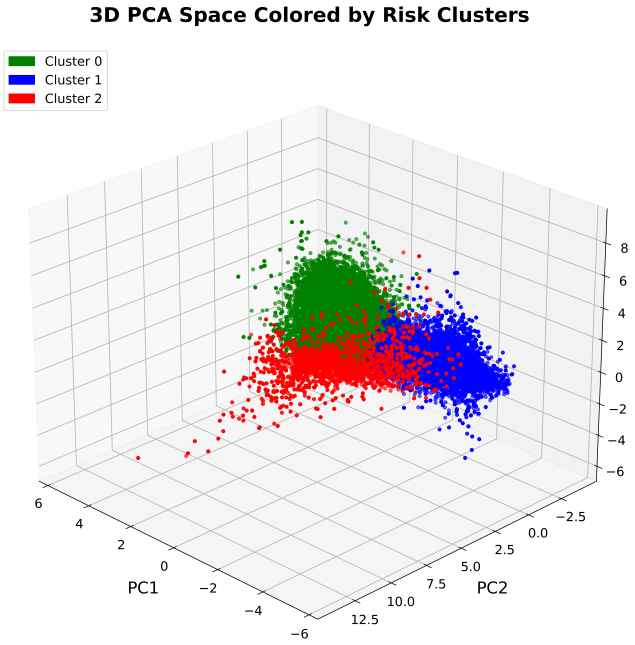

Fig. 4: 3D visualization of K-Means clustering applied to PCA-transformed meteorological data. Cluster 2 (in red) shows a high concentration of severe/extreme turbulence reports and was labeled as high-risk.

4.5 Supervised Model Training

A set of supervised classifiers was evaluated on the processed dataset using stratified 10-fold cross-validation. Models included:

—**XGBoost**, **Random Forest**, **LightGBM**, and **CatBoost** — ensemble tree-based methods known for strong performance on tabular data.

—**TabNet** — a neural network architecture tailored for tabular inputs with built-in feature selection.

—**SVM**, **KNN**, and **Naive Bayes** — lightweight statistical baselines for comparative analysis.

All models were trained using the same input feature set: PCA components, categorical encodings (Season, Aircraft, ARTCC), and risk cluster label. A classification threshold of 0.45 was used to favor recall for the minority class. Hyperparameters were selected using grid search or built-in optimization routines.

4.6 Evaluation Metrics

Model performance was evaluated using standard metrics:

—**Accuracy** — overall correctness across both classes.

—**Precision, Recall, F1-score** — with emphasis on Class 1 (severe/extreme).

—**ROC-AUC** — area under the receiver operating characteristic curve.

Confusion matrices and ROC curves were analyzed to assess trends in false positives and false negatives. The top-performing model (XGBoost) was selected for downstream case study evaluation in Section 5.

5. RESULTS AND DISCUSSION

The results of turbulence classification experiments were evaluated across multiple models and preprocessing settings using stratified 10-fold cross-validation, with special emphasis on detecting severe and extreme turbulence events (Class 1).

5.1 Model Performance Comparison

Figure 5 compares precision, recall, F1-score, and ROC-AUC across all classifiers. Among them, **XGBoost** consistently outperformed other models in all metrics, achieving:

—**Recall:** 0.91

—**F1-score:** 0.88

—**ROC-AUC:** 0.97

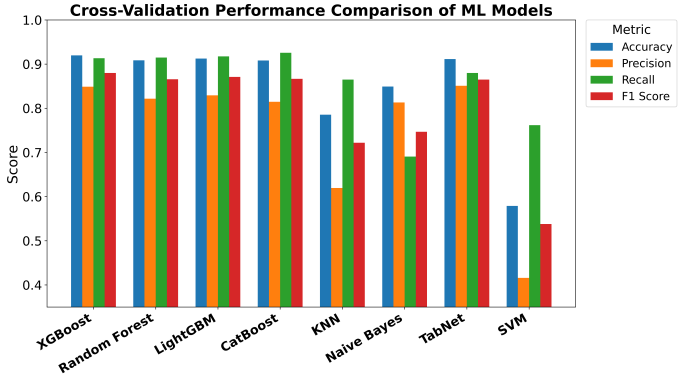

Fig. 5: Comparison of accuracy, precision, recall, and F1-score across all classifiers using 10-fold cross-validation. XGBoost achieved the best overall balance of sensitivity and precision.

Tree-based models (LightGBM, CatBoost, Random Forest) also performed strongly, particularly in recall and AUC, confirming their ability to capture nonlinear relationships in meteorological features. TabNet matched ensemble models in recall but showed higher fold-to-fold variability.

In contrast, baseline classifiers such as Naive Bayes and KNN achieved moderate accuracy but lagged in precision. While they managed decent recall, especially KNN, their tendency to overpredict turbulence led to more false positives, reducing their F1 scores. This suggests that while they captured some patterns, they lacked the specificity of the stronger ensemble models. On the other hand, SVM achieved moderate recall, but precision was significantly lower. Given the high-dimensional nature of the feature space and the size of the dataset, SVM's structure likely limited its ability to scale and adapt effectively in this context. While Mizuno et al. [12] previously applied SVM to smaller localized turbulence data, it did not generalize well here.

As shown in Figure 6, XGBoost, Random Forest, and LightGBM achieved the highest AUC scores, indicating strong class separability between turbulent and non-turbulent conditions. TabNet also reached competitive AUC values, though with greater variance across folds.

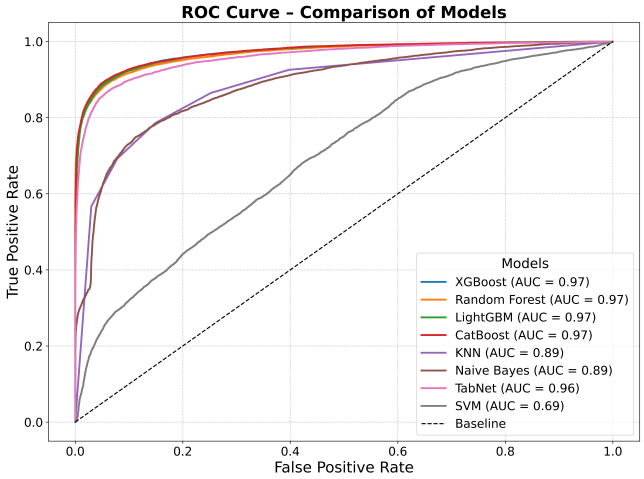

Fig. 6: ROC curves for all models. XGBoost, Random Forest, and LightGBM achieved the highest AUC values, indicating strong class separability.

5.2 Effect of Preprocessing Techniques

To isolate the contribution of each preprocessing component, an ablation study was conducted using XGBoost. Table 2 summarizes model performance under different combinations of class balancing and clustering. Several key findings emerged:

—Training on the raw dataset yielded high overall accuracy but poor recall (0.47), confirming that imbalance suppression is essential for detecting rare turbulence cases.

—Downsampling alone improved both precision (0.81) and recall (0.76), demonstrating that anomaly-aware selection of negative samples helps mitigate dataset skew.

—SMOTE alone raised recall (0.85) by generating synthetic high-risk samples, but this came at the cost of increased false positives and reduced precision.

—The right balance was achieved by combining SMOTE with anomaly-aware downsampling, which produced the highest F1-score (0.88) and balanced sensitivity with specificity.

—Incorporating PCA and K-Means clustering improved interpretability and enabled visualization of latent risk structures in the meteorological feature space. Although this step caused a minor decrease in precision, it enhanced the ability to conduct spatiotemporal analysis of high-risk clusters.

—The full pipeline, integrating all preprocessing steps, consistently achieved robust performance across folds, with a recall of 0.91 and an F1-score of 0.88, illustrating both reliability and generalization.

These findings validate the multi-stage pipeline design and highlight that combining both data-level balancing (SMOTE, downsampling) and model-level enhancements (PCA, K-Means clustering) provides strong predictive accuracy while preserving interpretability for operational use.

5.3 Comparison with Prior Work

While this study builds upon the methodology introduced by Mizuno et al. [12], the datasets and evaluation settings differ significantly. Mizuno et al. evaluated their approach on Quick Access Recorder (QAR) reports collected near Matsumoto Airport during early 2019, whereas this study applied the same methodology to more than 550,000 PIREPs combined with ERA5 data across the entire U.S. airspace from 2024.

Both studies employed a similar structure involving PCA for dimensionality reduction, K-Means clustering for turbulence-risk labeling, followed by Support Vector Machine (SVM) classification. Mizuno et al. [12] reported high accuracy on their localized dataset, with 87.5% recall for turbulent cases and only 0.7% false positives. However, when the same structure was applied in this study, the SVM model achieved 77.1% recall but misclassified nearly half of the negative cases, yielding just 51.9% accuracy on the negative class. This indicates sensitivity to dataset scale and diversity, with a high false alarm rate when applied across broader U.S. conditions. Figure 7 shows the contrast between Mizuno et al.'s [12] results and this study's SVM implementation.

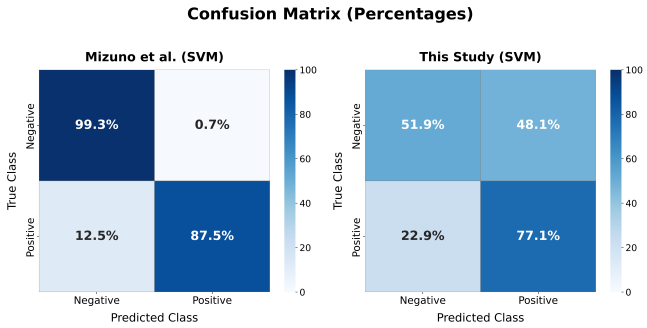

Fig. 7: Confusion matrix comparison between Mizuno et al.'s SVM model (left) and the SVM implementation in this study (right). The U.S.-scale dataset resulted in weaker separation between turbulent and non-turbulent cases, leading to a higher false alarm rate.

Table 2. : Ablation study of preprocessing steps on turbulence classification using XGBoost.

Preprocessing Method	Accuracy	Precision	Recall	F1-score
Raw data (no balancing)	**0.9663**	0.66	0.47	0.55
Downsampling only	0.8302	0.81	0.76	0.79
SMOTE only	0.9533	0.66	0.85	0.74
SMOTE + Downsampling	0.9218	**0.86**	0.90	**0.88**
PCA + KMeans	0.9275	0.53	0.88	0.66
Full Pipeline (CV)	0.9197	0.85	**0.91**	**0.88**

Note: Precision, recall, and F1-score are reported for the severe-extreme (positive) class, while accuracy reflects overall classification performance.

In contrast, the XGBoost model demonstrated strong classification performance across the same setting. It reduced the false negative rate to 9.3%, achieving 90.7% recall for severe–extreme turbulence and 93.6% accuracy on the negative class. Figure 8 presents a side-by-side comparison between Mizuno et al.'s SVM and the XGBoost model used in this study. These results highlight the scalability and robustness of the open-data-based pipeline, which maintains strong detection performance across diverse atmospheric conditions while avoiding the high false alarm rate of SVM.

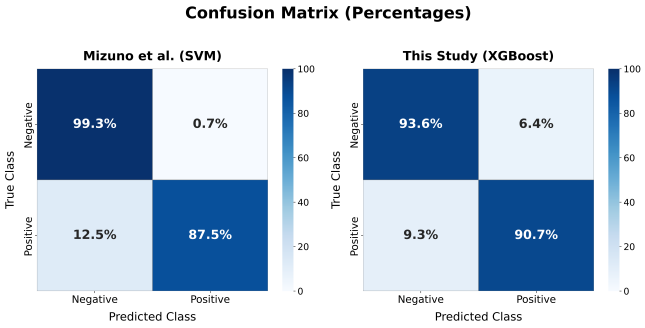

Fig. 8: Comparison of confusion matrices between Mizuno et al.'s SVM-based model (left) and the XGBoost model used in this study (right). The proposed pipeline shows improved recall and fewer false negatives across a broader and more diverse U.S. dataset.

5.4 Case Study: February 16, 2025

To assess real-world applicability, the trained XGBoost model was tested on an unseen, high-turbulence day: February 16, 2025. Between 13:00 and 17:00 UTC, 77 turbulent events were recorded across the eastern U.S. The model correctly classified 85 of 90 turbulence cases (true positives) and 93 of 102 negative cases (true negatives), with only 5 false negatives and 9 false positives. This strong performance, particularly under high-risk conditions, supports the model's utility in operational forecasting.

February 16 was selected because it concentrated one of the highest numbers of SEV–EXTRM reports between January and March 2025, providing a challenging yet representative test case. In addition to classification accuracy, the model demonstrated geographic consistency by showing that clusters of high-risk predictions overlapped with actual pilot reports, illustrating alignment between modeled outcomes and observed atmospheric instability.

Two high-risk identifications were analyzed in detail, one over the Ohio region and the other over North Carolina. Both showed consistent deviations in wind speed, vertical velocity, and humidity from 2024 seasonal baselines, reinforcing that the model captured meaningful atmospheric shifts. Table 3 summarizes the deviations for these two high-risk predictions during the event.

Figure 9 shows the spatial distribution of model predictions over the eastern U.S. during a high-risk window on February 16, 2025, where predicted turbulence clusters aligned closely with observed pilot reports.

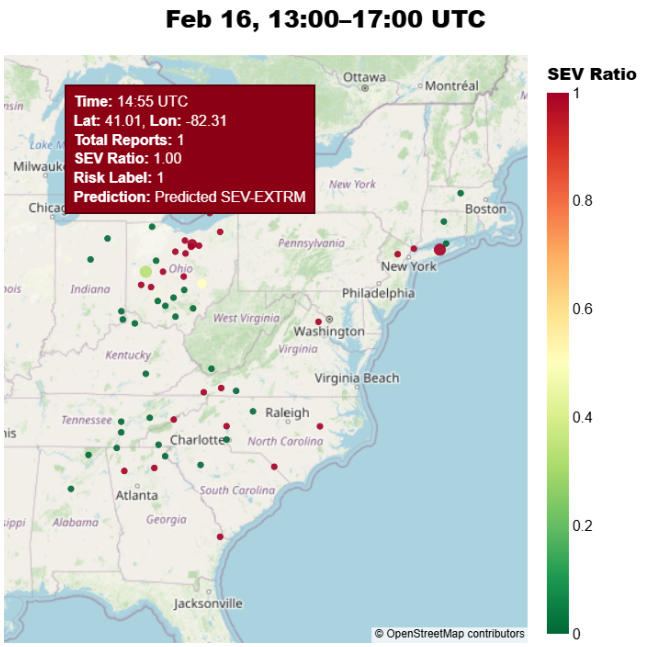

Fig. 9: Classified turbulence risk on February 16, 2025 (13:00–17:00 UTC), using the trained XGBoost model. Clusters of high-risk identifications align with observed pilot reports across the eastern U.S. Map data © OpenStreetMap contributors, available under the Open Database License (ODbL).

5.5 Interpretability and Deployment Potential

The use of PCA and K-Means clustering enables both interpretability and geographic visualization of turbulence-prone zones. The resulting model can be embedded into early-warning tools or route optimization software using only open-source inputs.

6. CONCLUSION AND FUTURE WORK

This study presents a scalable and interpretable machine learning pipeline for turbulence-risk identification using open flight and weather data. By combining over 550,000 PIREP reports with ERA5 reanalysis weather variables, a large, labeled dataset spanning the full U.S. airspace was constructed. The pipeline integrates anomaly-aware downsampling, SMOTE-based class balancing, PCA for dimensionality reduction, and K-Means clustering for latent risk discovery. Among all models evaluated, XGBoost achieved the best balance of precision and recall, with a recall of 0.91 and an F1-score of 0.88 across 10-fold cross-validation.

A case study on a real high-turbulence day in 2025 demonstrated the model's ability to detect severe turbulence conditions in operational settings with minimal false alarms. Compared to earlier approaches that relied on proprietary QAR data or localized models, the proposed method uses fully open data sources and generalizes across altitudes, aircraft types, and geographic sectors.

Several directions remain for further exploration:

—**Spatiotemporal modeling**: Incorporating time-aware models (e.g., LSTM, ConvLSTM) may improve detection of evolving turbulence patterns [5].

Table 3. : Meteorological deviations for two high-risk identifications on February 16, 2025, compared to 2024 seasonal averages. Wind speed and vertical velocity showed consistent elevation near severe turbulence events.

Date and Time (UTC)	Region (location)	Altitude (hPa)	Feature Deviations	Inferred Cause
Feb 16, 2025, 14:55	Ohio region (41.01N, -82.31W)	366.5 (27,500 feet approx.)	Wind speed increased by 6.5 m/s; vertical velocity +0.02 Pa/s; temperature 0.45°C lower; relative humidity 8.7% lower; wind direction shifted by -12°	Moderate vertical motion observed with stronger northward winds (v-component +14 m/s). Temperature was slightly lower than average SEV cases. Wind shear was slightly weaker.
Feb 16, 2025, 19:00	North Carolina region (35.54N, -79.67W)	548.9 (16,000 feet approx.)	Wind speed increased by 12.2 m/s; vertical velocity +0.50 Pa/s; temperature 3.1°C higher; relative humidity 29.6% lower; u-component of wind +14.6 m/s	Higher wind shear and warmer temperatures than the seasonal average. Low humidity and strong west–east winds likely contributed to localized turbulence formation.

Altitude values in parentheses are approximate conversions from pressure levels in hectopascals (hPa) to feet. All deviations are measured relative to 2024 seasonal baseline values.

—**Integration of trajectory data**: Combining turbulence reports with ADS-B or flight path data could enable predictive routing and real-time risk estimation.

—**Cross-dataset evaluation**: Extending evaluation across multiple years and additional atmospheric datasets (e.g., comparing ERA5 with GFS) would further strengthen confidence in model generalization and operational reliability.

—**Expert collaboration**: Involving meteorologists and aviation professionals could improve label reliability and validate feature importances.

—**Interactive visualization**: Coupling the model with map-based tools would help forecast users explore and explain high-risk zones using interpretable clusters.

By demonstrating that turbulence identification is feasible using only open data and reproducible methods, this work lays the foundation for future integration into safety tools, forecasting systems, or decision-support software for both pilots and air traffic controllers.

7. REFERENCES

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[2] Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[5] Mounir Chrit and Marwa Majdi. Operational wind and turbulence nowcasting capability for advanced air mobility. *Neural Computing and Applications*, 36(18):10637–10654, 2024.

[6] Ivan Bitar Fiuza de Mello, Gutemberg Borges França, and Haroldo Fraga de Campos Velho. Enhancing clear air turbulence prediction: A comparative analysis of machine learning algorithms using gfs forecast and era-5 reanalysis data. 2024.

[7] Federal Aviation Administration. Turbulence: Staying safe. `https://www.faa.gov/travelers/fly_safe/turbulence`.

[8] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[9] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

[10] Afaq Khattak, Jianping Zhang, Pak-Wai Chan, Feng Chen, and Abdulrazak H Almaliki. Aviation safety at the brink: Unveiling the hidden dangers of wind-shear-related aircraft-missed approaches. *Aerospace*, 12(2):126, 2025.

[11] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[12] Shinya Mizuno, Haruka Ohba, and Koji Ito. Machine learning-based turbulence-risk prediction method for the safe operation of aircrafts. *Journal of Big Data*, 9(1):29, 2022.

[13] Domingo Muñoz-Esparza and Robert Sharman. An improved algorithm for low-level turbulence forecasting. *Journal of Applied Meteorology and Climatology*, 57(6):1249–1263, 2018.

[14] National Transportation Safety Board. Preventing turbulence-related injuries in air carrier operations conducted

under title 14 code of federal regulations part 121. Technical Report SS-21/01, NTSB, 2021. https://www.ntsb.gov/safety/safety-studies/Documents/SS2101.pdf.

[15] Robert Sharman, C Tebaldi, G Wiener, and J Wolff. An integrated approach to mid-and upper-level turbulence forecasting. *Weather and forecasting*, 21(3):268–287, 2006.

[16] Luke N Storer, Paul D Williams, and Philip G Gill. Aviation turbulence: dynamics, forecasting, and response to climate change. *Pure and Applied Geophysics*, 176:2081–2095, 2019.

[17] Paul D Williams. Increased light, moderate, and severe clear-air turbulence in response to climate change. *Advances in atmospheric sciences*, 34(5):576–586, 2017.