Multiclass Flight Anomaly Detection Using Sensor Fusion Based on Dempster-Shafer Theory

Ezequiel Juarez Garcia*, Szilard L. Beres†, and Markus L. Mulvihill‡
University of Florida, Gainesville, FL, 32611

Chad L. Stephens§
NASA Langley Research Center, Hampton, VA 23681

Nicholas J. Napoli¶‖
University of Florida, Gainesville, FL, 32611

As aviation systems in commercial operations continue to grow in complexity, the anomalies exhibited by these systems become more elaborate and difficult to detect. To address the challenge of detecting these complex anomalies, deep learning models have been used extensively in aviation anomaly detection studies, at the expense of end-user interpretability. Aiming to maintain the same level of interpretability as traditional threshold-exceedance methods, we continue our development of prediction models using ordinal patterns and their distributions throughout the flight. Specifically, this study extends our work into multiclass anomaly detection using sensor fusion based on Dempster-Shafer theory (DST), a second-order probability theory used to combine information from different sources of evidence. Our approach uses DST to reduce the uncertainty in the class predictions of an ensemble of classifiers. These classifiers rely on the similarity between flight data and class templates to make a prediction of the state of the aircraft. Our approach aims to take advantage of simple models trained on interpretable features (ordinal patterns) to correctly predict an anomaly and identify the flight dynamics linked to the anomaly. Our results show an improvement when using DST-based sensor fusion over simple majority voting. Additionally, our results provide insight into aircraft states linked to rare high-risk anomalies.

I. Introduction

Modern safety management programs seek to identify the root causes of aviation accidents using tools and data currently available to continue improving the safety of commercial air transportation and safeguard against future risks. As the national airspace system (NAS) continues to expand and prepare for increasing air traffic and the introduction of unmanned aircraft systems (UAS)[1], more research into anomaly detection and prediction is necessary to maintain and improve airspace resilience and safety. In this study, we refer to anomalies as the events leading up to safety incidents (i.e., precursor events), but they can also refer to the safety incidents and accidents themselves. In addition to achieving high performance, anomaly detection models should be fundamentally capable of translating to future aircraft and retain the end-user (e.g., pilot) interpretability of simpler traditional methods. Therefore, the design of new anomaly detection paradigms needs to factor in model complexity to achieve these goals. Models with high complexity tend to result in lower model interpretability and vice versa. The move away from traditional threshold-exceedance and safety bound check anomaly detection methods has resulted in the proliferation of complex machine learning (ML) and, in particular, deep learning (DL) models to identify anomalies in aviation operations [2, 3]. While DL models may provide a seemingly one-stop solution to nearly all aviation safety problems, these models are not necessarily easy to interpret or

*Graduate Research Assistant in the Human Informatics and Predictive Performance Optimization (HIPPO) Laboratory, Department of Electrical and Computer Engineering, AIAA Student Member (e-mail: ejuarezgarcia@ufl.edu).

†Graduate Research Assistant in the Human Informatics and Predictive Performance Optimization (HIPPO) Laboratory, Department of Electrical and Computer Engineering (e-mail: szilard.beres@ufl.edu).

‡Undergraduate Researcher in the Human Informatics and Predictive Performance Optimization (HIPPO) Laboratory, Department of Electrical and Computer Engineering (e-mail: markus.mulvihill@ufl.edu).

§Researcher, Crew Systems & Aviation Operations Branch and System-Wide Safety Project (e-mail: chad.l.stephens@nasa.gov).

¶Assistant Professor and Director of the Human Informatics and Predictive Performance Optimization (HIPPO) Laboratory, Department of Electrical and Computer Engineering, AIAA Senior Member (e-mail: n.napoli@ufl.edu).

generalize to other airframes. To address the interpretability of anomaly detection, this study provides an approach to anomaly detection using simple detection models that do not rely on sophisticated machine learning models.

The features used to train the anomaly detection models need to be interpretable themselves and extracted from datasets in an informed manner to increase the overall interpretability of the model. Informed featured extraction means that the features extracted from the data maintain a close link to real-life quantities or processes. Albeit requiring some level of domain-expert knowledge, the use of informed feature engineering allows for the creation of anomaly detection paradigms that do not need to rely on complex machine learning models for feature extraction or anomaly detection. By carefully extracting important, interpretable features from the multitude of sensors in modern aircraft, our goal is to obtain nearly the same level of performance as DL anomaly detection models. Continuing our previous work on interpretable anomaly detection [4], we rely on the orderings of multiple signals to act as the interpretable features behind our prediction models. Figure 1 provides a simple illustration of these orderings, formally known as *ordinal patterns*. A key insight in our new methodology is to harness the prediction capability of specific sensors, some of which may be better suited for detecting certain types of anomalies. For example, airspeed and descent rate may be better predictors of instability during approach than altitude. By extending this concept to a multivariate case, the ordinal patterns of multiple signals can provide unique signatures of an anomaly prior to its offset. Therefore, this work continues the use of ordinal patterns obtained from combinations of multiple sensors. We rely on sensor fusion to discern the ordinal patterns (i.e., ordinal patterns) driving the anomaly. The use of a multitude of small and simple prediction models, combined with sensor fusion, will allow us to maintain a higher level of interpretability compared to more complex deep learning methods.

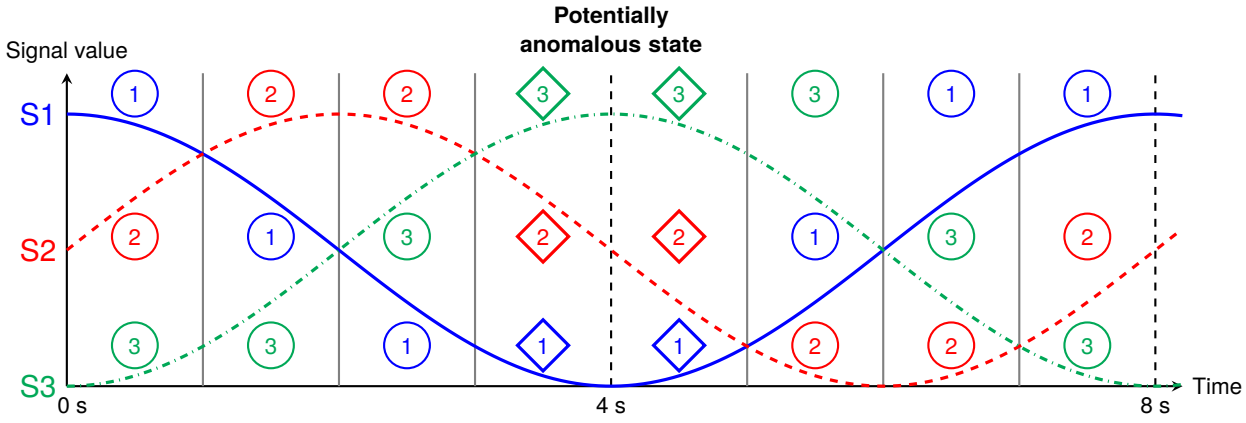


Fig. 1 Explanatory figure on ordinal patterns using 3 different signals (S1, S2, and S3). Ordinal patterns provide a simple approach to modeling the multivariate interactions of flight variables by looking at how signals are ranked, or ordered, at any point in time. The specific variable interactions between 3 and 5 seconds (i.e., the ordinal pattern (1, 2, 3) when S1 $\leq$ S2 $\leq$ S3), may indicate the current or future presence of an anomaly or its precursor. Anomalous and nominal ordinal patterns act as the interpretable features used to train the prediction models.

Prior Work. Traditional anomaly detection relies on predefined thresholds or safety bounds to register whether an anomaly has occurred. To keep up with the growing complexity of aviation systems [5], sophisticated deep learning approaches such as [2, 3, 6] have been introduced to identify anomalies during and prior to their onset. These approaches rely on a one-model-fits-all strategy to achieve high classification performance. This performance comes with the trade-offs of large training times and reduced modularity and generalizability. A less common approach to anomaly detection is to rely on smaller, simpler models. While achieving lower performance compared to complex deep learning models, a collection of small models has a higher degree of modularity, since individual models can be readily replaced, added, or removed from the prediction framework to cater to the aircraft's sensors. To combine or *fuse* the prediction of multiple sensors (i.e., prediction models), a simple strategy such as averaging can be used. More advanced fusion methods, such as [7], utilize more mathematically rigorous algorithms to resolve conflicting model predictions and disagreement between models. These sensor fusion methods are able to "weigh" the predictions of certain models higher for certain anomalies if the models are strong predictors of the anomaly. To build a sensor fusion framework for anomaly detection, our previous work focused on creating simple and interpretable prediction models [4, 8], a necessary

step in creating the fusion framework. Therein, we created naive Bayes classifiers to detect the presence of an unstable approach. An approach is considered unstable when it exhibits significant lateral, vertical, or speed deviations from a predefined flight path [9, 10]. Our continuing work in interpretable anomaly detection presents an opportunity to use both interpretable models and sensor fusion to perform multiclass detection of high-risk anomalies during approach.

In-flight anomaly detection requires the correct prediction of the anomaly, as well as the proper identification of the events or variables driving the anomaly. The key events occurring prior to the onset of an anomaly are known as *precursors*. A rising trend in aviation anomaly detection over recent years has been the identification of precursors to predict anomalies ahead-of-time. Precursor identification aims to give the end-users, such as pilots and air traffic services, adequate time to take corrective actions to prevent or reduce the effects of an anomaly. To tackle this complex task, deep learning models are typically utilized to perform precursor mining [11]. Despite progress in increasing the interpretability of deep learning models [3], current methods do not reach the same level of interpretability as traditional threshold-exceedance methods. Our prior work has demonstrated the viability of using ordinal patterns—a method for summarizing the ordering of multiple signals—as interpretable features to build prediction models [4]. However, two key limitations of our previous work were its binary class design (stable vs. unstable approach) and its inability to link the underlying flight dynamics to aircraft instability. The latter limitation is also present in other literary works in the field. Information on the flight dynamics at the root of an anomaly can help improve pilot training and aid corrective action. Consequently, the simple and interpretable anomaly detection method provided in this study can help identify the flight dynamics linked to in-flight anomalies during approach and landing to aid in risk remediation.

Challenges. The rise in the complexity of models for detecting in-flight anomalies has mainly stemmed from the inadequacy of traditional threshold-exceedance methods at detecting elaborate patterns of failure or anomalies with high accuracy. Moreover, certain anomalies may even go undetected when solely relying on traditional methods [12]. Recent accidents such as the Lion Air catastrophe caused by a design flaw in the Boeing 737 MAX's Maneuvering Characteristics Augmentation System (MCAS) [13] highlight the need for more robust methods of detecting complex anomalies. Capturing more nuanced anomalous patterns and relationships between flight dynamics has given way to the use of deep learning. However, despite their wide success, the penalty incurred by deep learning models is the poor interpretability of their inner workings and sometimes their output. Additionally, even if interpretable features are used to train a deep learning model (e.g., mean values of flight variables), this does not necessarily imply that the features learned by the model will maintain the same level of interpretability. Thus, despite the thousands of features that typical deep learning models are trained with, it can be extremely difficult to extract a single interpretable statistic or feature from inside the model to present to the end-user as the explanation of an anomaly. The interpretability of the system must not only be understandable to the researcher or practitioner developing the detection model but, more importantly, to the pilots and air traffic services that will ultimately interact with the model once it is implemented in aviation systems.

Insights. Adding more sophisticated features and training algorithms to detection models naturally results in more complex models. Going against the trend of relying on deep learning, we instead seek to create an *ensemble* (i.e., collection) of smaller and simpler prediction models that collectively provide nearly the same level of performance as more sophisticated approaches. This approach borrows from the field of sensor fusion, where a combination of different sensors or models can provide a better estimate of unknown state [14]. Sensor fusion addresses the issue of poorly trained models that will inevitably be present in the model ensemble. Despite the varying uncertainty in the anomaly prediction from each sensor, sensor fusion allows us to combine disparate sources of information to reduce the uncertainty of the final prediction. Additionally, the interpretability of the features used to train the prediction models also plays an important role in anomaly detection. Combining simple models with interpretable features allows us to gain insight into what is causing an anomaly. Our previous work in more interpretable features capable of predicting in-flight anomalies [4, 8] culminated in our introduction of ordinal patterns (see Fig. 1 for description). With ordinal patterns, we were able to predict unstable approaches 1 minute ahead-of-time with an accuracy of 0.69 and a recall of 0.73 and 30 seconds ahead with an accuracy of 0.70 and a recall of 0.86. This level of performance was achieved using only 4 sensors and a single binary classifier. This study continues our use of ordinal patterns with more sensors to help us identify multiple anomalies during approach and landing. We use Dempster-Shafer theory (DST) [15], also known as evidence theory, to combine multiple model predictions. DST is a probabilistic theory capable of modeling uncertainty in the true state of a system by assigning a mass to any combination of states. The concept of uncertainty provided by DST allows us to work with multiple prediction models that may not be in agreement on what flight dynamics are the strongest predictors of an anomaly. For this type of analysis, our fusion methodology relies on a balanced testing set to reduce the negative effects of class imbalance on the results. With this information, we are not only able to use sensor fusion to predict anomalies prior to their onset, but also to identify the flight dynamics linked to anomalies. This can lead to improved pilot training and risk mitigation procedures that can help reduce the occurrence of in-flight anomalies.

Contributions. In this study, we demonstrate the performance of an interpretable multiclass anomaly detection method capable of detecting four types of anomalous classes (including nominal) during approach and landing. Our approach uses DST to combine the uncertain predictions of various models. More specifically, our contributions in this study are to

- Establish the viability of small classifiers trained on reduced datasets for predicting in-flight anomalies.
- Resolve uncertainty in the predictions of multiple classifiers using DST to obtain a more accurate final anomaly prediction.
- Identify the flight dynamics embedded within ordinal patterns linked to specific anomalies.

The above goals will allow us to gain insight into what variable interactions are most important in identifying and preventing anomalies. This analysis can help us better understand high-risk states of current and future autonomous aircraft and provide safety recommendations to safety stakeholders such as manufacturers, aviation agencies, and airlines. Details of our proposed methodology are provided in the following section.

II. Methodology

Our previous work revealed that certain ordinal patterns were more likely to occur during unstable approaches compared to stable approaches [4]. In that study, we examined the two-class approach instability problem using the ordinal patterns of flight variables linked to unstable approaches: airspeed, glideslope deviation, localizer deviation, and vertical velocity [16]. This present study continues the use of ordinal patterns as interpretable features that can be used to create anomaly prediction models. These ordinal patterns are created from dozens of combinations of flight variables (e.g., airspeed, glideslope deviation, etc.) listed in Table 3. This marks an improvement over our previous methodology, which only used a single combination of variables. We also extend our work from binary classification into a 4-class problem using sensor fusion. More specifically, we rely on the distributions of ordinal patterns at each time step during the flight to make a prediction of an anomaly. It is especially important that we detect these anomalies prior to the 1,000 ft height above ground level (AGL) threshold as, under instrument flight rules (IFR), an aircraft must be stabilized prior this height AGL threshold. Moreover, using DST as the mathematical foundation for our fusion framework, we can reduce the uncertainty in the information provided by the pattern distributions to improve the final anomaly prediction. Other improvements to our methodology are detailed in the following sections. Before proceeding, we summarize the steps in our methodology in the following high-level overview:

1. Given all the flight data, normalize the flight variables to a common value range
2. Create the training and test data splits of normalized data using 5-fold cross validation [17]
3. With the training data:
   a) Create reduced datasets (i.e., subsets) of the normalized data by randomly sampling 4 flight variables from a total of 20
   b) Extract the ordinal patterns in each 4-variable subset
   c) Calculate the likelihood distributions of the ordinal patterns
4. With the test data:
   a) Find the similarity between the ordinal patterns in a flight and the ordinal patterns in likelihood distribution
   b) Use sensor fusion to reduce the uncertainty in the information provided by the likelihoods and improve prediction

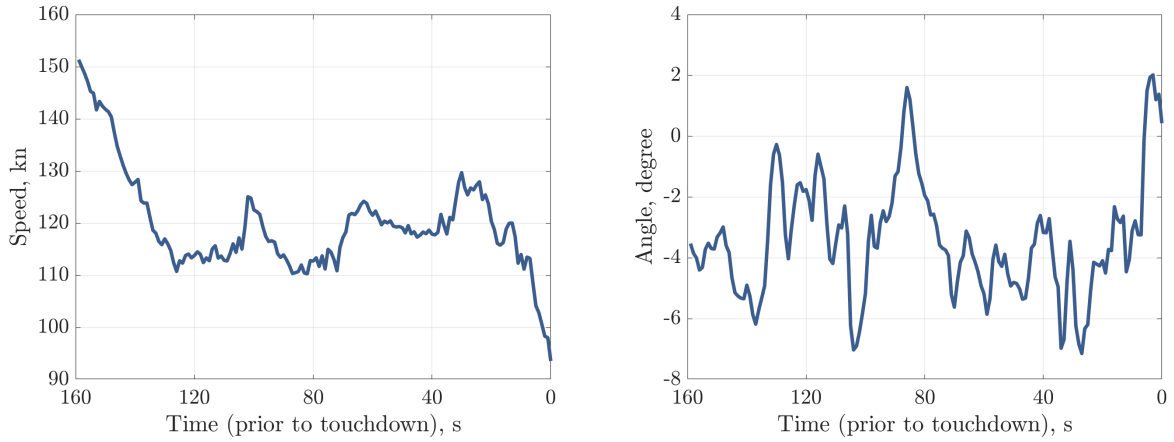
A. Dataset Overview and Normalization

The dataset used in this work was developed for the multiclass anomaly detection study in [18]. We refer to this dataset as the *curated flight dataset* or, simply, as the *flight data* in this study. This dataset contains approximately 100,000 flight records, with each record containing the time series of 20 flight variables. The complete list of flight variables in the dataset can be found in Appendix V.A. Each variable time series captures a 160-second long data window recorded prior to touchdown. A uniform sampling rate of 1 Hz is used in all time series. The dataset contains four classes: *nominal, high speed, high path,* and *late flaps setting*—the last three of which are considered the anomalous classes. The thresholds used to calculate these anomalies are summarized in Table 1. The flight data is derived from a larger public dataset named the NASA DASHlink Sample Flight Data [19]. This larger dataset contains more flight records, 180,000 in total, and more flight variables. Compared to the original, the curated flight data contains a subset of data with labels obtained with the aid of subject-matter experts [18]. Figure 2 provides examples of time series found in the flight data.

As shown in Figure 1, ordinal patterns can be used to capture the multivariate state of an aircraft's dynamics. Before

Table 1 Thresholds used to define the anomalous classes in this study. Flights that did not exhibit any of the anomalies below where categorized as nominal.

Anomaly	Threshold
High speed	Difference in CAS - CASS 2σ above nominal
High path	Height AGL 2σ above nominal at same point in time
Late flaps setting	Flaps deployed 60 seconds later than nominal



(a) Airspeed (b) Pitch angle

Fig. 2 Examples of time series found in the flight data. The two flight signals are from the same flight record.

extracting the ordinal patterns from the multivariate time series of each flight, we must first normalize the flight variables to a common value range. The normalization used in this study maps the $\mu \pm 2\sigma$ (two standard deviations, SDs, about the mean) range of each flight variable to the closed interval $[0, 1]$. This normalization technique assumes that the variable distributions at each time step are normally distributed. The use of 2 SDs for normalization reduces the effects of outliers on the usable normalized range. Moreover, due to the evolution of the time series, we chose the largest 2 SD bounds during the 160 seconds of flight data to correspond to the values 0 and 1. Naturally, values outside the 2 SD bounds will be mapped outside the range $[0, 1]$. Based on Figure 3, data past 1,000 ft height AGL—about 90 seconds prior to touchdown—is not analyzed in this study. After normalization, the example time series in Figure 2 become those shown in Figure 4.

B. Data Partitioning and Approximating Likelihood Distributions

1. Partitioning the Flight Data

Within each fold, the normalized data is partitioned to create reduced datasets, or subsets, according to different combinations of flight variables. These data subsets are comprised of the data of 4 flight variables randomly chosen from the pool of 20 flight variables (see Appendix V.A for description of all flight variables). A total of 20, 4-variable combinations are created for each data fold. The importance of using multiple combinations of flight variables (i.e., creating the reduced datasets) will become more apparent later on, when we use sensor fusion to combine the information provided by each variable combination. The process of creating the data subsets is visualized in Figure 5. To understand this figure and the rest of this text, we denote the number of samples in a flight variable time series as T; the number of flight variables in the flight data as N; the number of flights in the data belonging to class c as M^c; and the number of

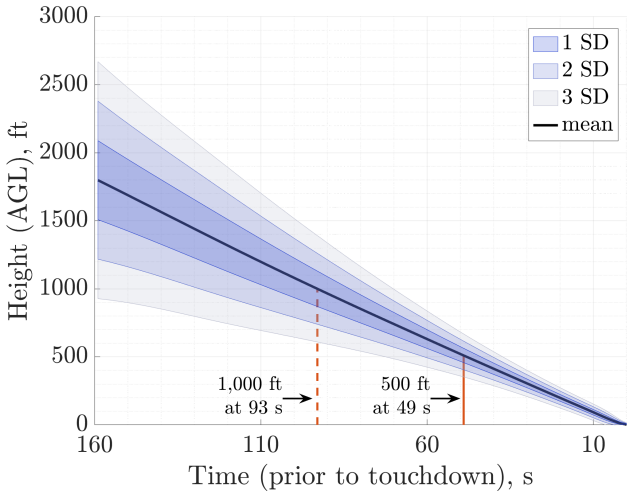
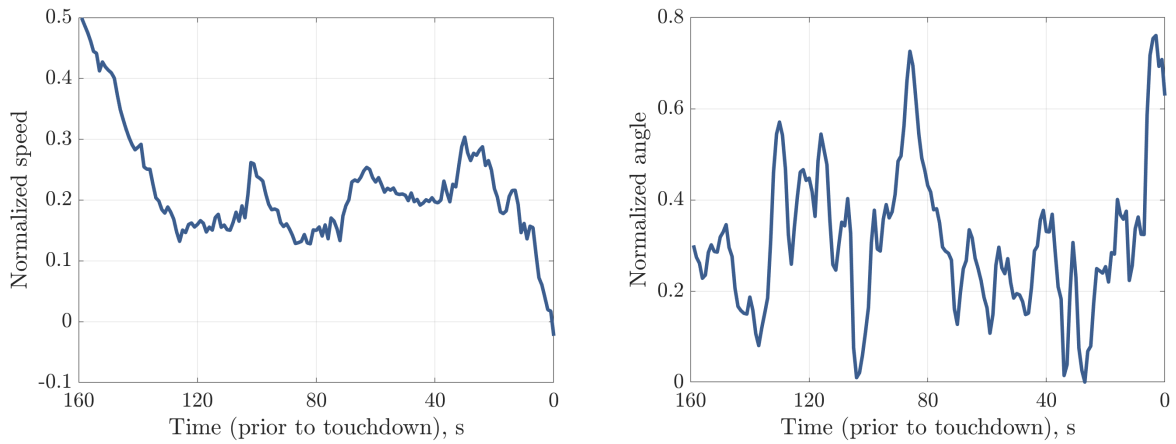


Fig. 3 Distribution of height AGL during runway approach across all 100,000 flights in flight data.



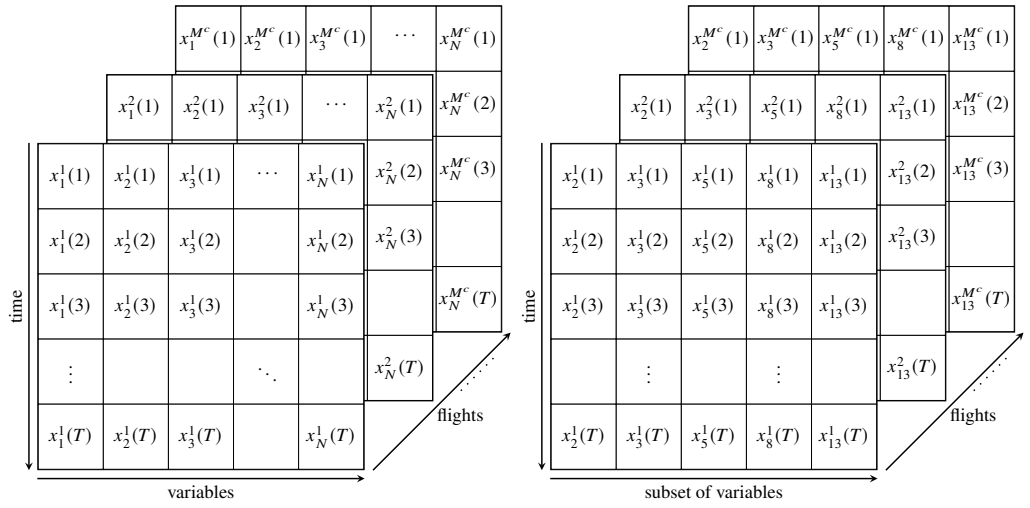
(a) Airspeed (b) Pitch angle

Fig. 4 Examples of normalized time series from Figure 2. The two flight signals are from the same flight record.

4-variable combinations as R. The normalized data $\mathcal{D}^c$ of class c (within a fold) is used to create the reduced datasets from which the ordinal patterns are extracted later on. Care was taken to prevent the creation of subsets where any two variables have a Pearson correlation coefficient greater than 0.3. This was done to discard variables that do not provide any new information.

2. Extracting Ordinal Patterns

Ordinal patterns capture the ordering or ranking of multiple signals (i.e., a multivariate time series). Given symbolic representations for each signal in a multivariate time series (e.g., integers 1, 2, . . .), each ordinal pattern stores the symbolic representation of the signals when sorted in increasing order. To illustrate this, we refer to the simple example in Figure 1. In the figure, the potentially anomalous state that occurs at around 4 seconds is equivalent to the ordinal pattern $\pi = (1, 2, 3)$. The symbols within the ordinal pattern π represent the signals S1, S2, and S3, respectively. The



(a) Normalized class data array $\mathcal{D}^c$

(b) Example of reduced dataset d_r^c

Fig. 5 Data structures used during the ordinal pattern template creation process. For each class c in a fold, the 3D array $\mathcal{D}^c$ in (a) contains the normalized time series data of the M^c flights belonging to that class. The number of time steps in a time series is denoted by T and the number of flight variables is denoted by N. Based on the 4-variable combinations generated for a fold, $\mathcal{D}^c$ is partitioned into subsets $\{d_r^c\}_{r=1}^R$. An example reduced dataset d_r^c is provided in (b).

entire evolution of the ordinal patterns in the example figure can be written in matrix notation as

$$\mathbf{\Pi} = \begin{bmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \\ \pi(4) \\ \pi(5) \\ \pi(6) \\ \pi(7) \\ \pi(8) \end{bmatrix} = \begin{bmatrix} 3 & 2 & 1 \\ 3 & 1 & 2 \\ 1 & 3 & 2 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \\ 2 & 1 & 3 \\ 2 & 3 & 1 \\ 3 & 2 & 1 \end{bmatrix},$$

where each row in the matrix represents a time step. The ordinal patterns created from the actual flight data are extracted from the normalized reduced datasets shown in Figure 5. The pattern extraction process follows the same steps as in the aforementioned example. The data produced at the end of this process is a set

$$\mathcal{F}_r^c = \left\{ \mathbf{\Pi}_r^c(1),\ \mathbf{\Pi}_r^c(2),\ \ldots,\ \mathbf{\Pi}_r^c(M^c) \right\} \tag{1}$$

for each combination of $r = 1, \ldots, R$ and $c \in \{\text{nominal}, \text{high speed}, \text{high path}, \text{late flaps}\}$. The index inside the parentheses denotes the number of a flight within a class. Similar to Figure 5, this set can be stored as a 3D array after computation. The pattern extraction process is repeated in the same manner for flights in the testing folds.

3. Creating Ordinal Pattern Likelihood Distributions

Given the ordinal patterns of hundreds of flights in a training fold, we use them to approximate a likelihood distribution of the ordinal patterns. That is, based on Bayes' rule, we would like to estimate the likelihood probability in

the formula

$$\overbrace{p(c_j|\pi_i)}^{\text{posterior}} = \frac{\overbrace{p(\pi_i|c_j)}^{\text{likelihood}}\overbrace{p(c_j)}^{\text{prior}}}{\underbrace{p(\pi_i)}_{\text{evidence}}}.$$

Here, $p(c_j|\pi_i) \coloneqq p(c = c_j|\pi_i)$ denotes a probability, whereas $p(c|\pi_i)$ would denote a probability distribution. The approximate likelihood distribution $p(\pi|c_j)$ (for a class $j = 1, 2, 3, 4$) is a probability mass function (PMF) of the ordinal patterns conditioned on class j. Due to a class imbalance, we randomly sampled 760 flights from each class to create the likelihood approximations. This number corresponded to the class with the lowest number of training samples in a fold, which was the late flaps setting class. The class imbalance is characterized by the following distribution of flights for the entire dataset (99,837 flights): 89.91% of flights are nominal, 7.02% are high speed, 2.21% are high path, and 0.96% are late flaps setting. In each of the $R = 20$, 4-variable combinations, a total of 24 ordinal patterns are possible. This creates a likelihood PMF with 24 probability values, one for each pattern. Specifically, at each time step, we approximate the likelihood at that time step by creating a normalized histogram per class of the number of flights that exhibited each pattern. Given a variable combination number r and class c, the resulting likelihood distributions are grouped into a set

$$P_r^c = \left\{p_{r,1}(\pi|c),\ p_{r,2}(\pi|c),\ \ldots,\ p_{r,T_{\text{stop}}}(\pi|c)\right\}, \tag{2}$$

where the second index in the subscript denotes the time step. Due to our focus on the flight data prior to 90 seconds before touchdown (see Figure 3), $T_{\text{stop}} = 70$. This means that we only look at the first 70 seconds in the 160-second long flight data to make predictions about the aircraft's future state, after it has crossed the 1,000 ft height AGL threshold.

C. Model Predictions and Sensor Fusion Framework

Due to the inseparable tie between the similarity of ordinal patterns and the fusion framework, the explanation of the model creation process was left until now. Two more items are in order, however, before we can explain the model creation process. First, a definition of ordinal pattern similarity and, second, some background on Dempster-Shafer theory (DST).

1. Background on Ordinal Pattern Similarity and DST

Defining Ordinal Pattern Similarity: Our work in [4] provides the motivation behind our definitions of distance and similarity between ordinal patterns. Here, we mainly restate the definition to understand the methodology. Given two equal-length ordinal patterns π_i and π_j, we define the distance between π_i and π_j using the formula

$$\overline{\text{dist}}(\pi_i, \pi_j) = \begin{cases} 0, & |\pi_i| = |\pi_j| = 0 \\ \overline{\text{dist}}\big(\text{tail}(\pi_i), \text{tail}(\pi_j)\big), & \pi_i(1) = \pi_j(1) \\ 1 + \overline{\text{dist}}\big([\text{tail}(\pi_i), \pi_i(1)], \pi_j\big), & \text{otherwise.} \end{cases} \tag{3}$$

In the equation above, $\pi(1)$ is the first element of π, $|\pi|$ is the length of π, $[\pi_1, \pi_2]$ is the concatenation of two patterns, and $\text{tail}(\pi)$ is the ordinal pattern excluding the first element. The distance metric computes the number of index shifts or moves required to make π_i identical to π_j, or vice versa. For example, the distance between $\pi_1 = (2, 3, 4, 1)$ and $\pi_2 = (1, 2, 3, 4)$ is 3, due to the three shifts to the left required to move the symbol 1 from the fourth position to the first position in π_1 to make it equal π_2. This simple example also highlights an important property of the calculation of $\overline{\text{dist}}(\pi_1, \pi_2)$, in that patterns do not wrap at the end to the beginning, meaning that the 1 in the last position of π_1 cannot be shifted to the right by one position to make it equal π_2.

Normalizing the value of Eq. (3) by its upper bound equal to $\binom{m}{2}$, where $m = |\pi_i| = |\pi_j|$, we define the normalized distance metric

$$\text{dist}(\pi_i, \pi_j) = \frac{\overline{\text{dist}}(\pi_i, \pi_j)}{\binom{m}{2}}. \tag{4}$$

With the normalized distance, we obtain the similarity between two ordinal patterns π_i and π_j using

$$\text{sim}(\pi_i, \pi_j) = 1 - \text{dist}(\pi_i, \pi_j). \tag{5}$$

Given m-length patterns, similarity can take on $\binom{m}{2} + 1$ values on the interval [0, 1], with a spacing of $1/\binom{m}{2}$ between successive values. In the previous example, $\text{sim}(\pi_i, \pi_j) = 0.5$. Any discussion of similarity between patterns in this study uses Eq. (5) to mathematically quantify the similarity.

Background on Dempster-Shafer Theory: DST is a generalization of probability theory used to quantify and reduce the uncertainty contained within sources of information. In this study, these sources of information are the likelihood distributions (i.e., PMFs) approximated by histograms. Due to factors such as heavy class imbalance, these distribution approximations have an inherent degree of uncertainty. Uncertainty plays a central role in DST, where it used to capture the degree of ignorance (i.e., lack of knowledge) in the state of a system. DST expresses mutually exclusive and exhaustive states of a system as *propositions*. The set of all propositions is known as the *frame of discernment* (FoD) Ω. In the context of classification, the framework of discernment consists of the nominal and anomalous classes. In other words, $\Omega = \{c_1, c_2, c_3, c_4\}$, where from here on forward, c_1 = nominal, c_2 = high speed, c_3 = high path, and c_4 = late flaps. All elements of the power set $\mathcal{P}(\Omega)$ are also treated as propositions. This allows the possibility to assign support to the classes themselves (single element sets, singletons) and also to combinations of classes (doubletons, tripletons, and so on). Support is assigned to each element in $\mathcal{P}(\Omega)$ with a *mass assignment* function $m : \mathcal{P}(\Omega) \longmapsto [0, 1]$. This mass assignment function must satisfy the following properties:

$$m(\emptyset) = 0; \quad \sum_{A \subseteq \mathcal{P}(\Omega)} m(A) = 1 \tag{6}$$

The support given to the FoD, Ω, models the level uncertainty in the information, where if $m(\Omega) = 1$, we are fully ignorant (i.e., have no knowledge) of the true state of the system given some information from a source. A source of information is referred to as a *body of evidence* (BoE) in DST.

2. Uncertainty Quantification and Model Predictions

A key element in a sensor fusion framework, especially one based in DST, is the manner in which the uncertainty of a sensor's information is quantified. The set of distributions P_r^c given in Eq. (2) are used to create a prediction model, or "sensor", $\mathcal{M}_r$. Formally, in DST, each model is a BoE triplet $\mathcal{M}_r = \{\Omega, F, m_r(\cdot)\}$, where F is the set of propositions in $\mathcal{P}(\Omega)$ that have non-zero mass. Due to our exclusive use singleton propositions and Ω, F consists of only 5 elements. To assign support to Ω (i.e., quantify the prediction uncertainty) and single propositions (i.e., classes), it is essential to define a valid mass assignment function $m_r(\cdot)$ for each BoE. We begin by defining the uncertainty in the prediction by using the similarity metric to measure the degree of agreement and disagreement between the observed pattern and other patterns, at any given time step.

First, let Δ_s be the interval/spacing between two consecutive similarity values. As discussed earlier for Eq. (5), given m-length patterns, the number of unique similarity values is $\binom{m}{2} + 1$; therefore, $\Delta_s = 1/\binom{m}{2}$. Given an observed pattern at time step t, $\pi^\star(t)$, and $K = \binom{m}{2}$ similarity values, let N_k (for $k = 0, \ldots, K-1$) be the number of patterns whose similarity to the observed pattern is $k\Delta_s$. Furthermore, given class $c^\star = \text{argmax}_{c_j}\, p(c_j|\pi^\star)$, let A_k be the proportion of patterns whose similarity is equal to $k\Delta_s$ and have $c^\star$ as their highest probability class. We refer to A_k as the agreement factor. With this, we define the uncertainty in the class agreement between ordinal patterns as

$$m(\Omega_A) = 1 - \frac{1}{b} \sum_{k=0}^{K-1} (k+1)\Delta_s A_k, \tag{7}$$

where b is a normalization constant equal to $\frac{\binom{m}{2}+1}{2}$ so that $0 \leq m(\Omega_A) \leq 1$. In the equation above, we expect similar patterns, in reference to $\pi^\star(t)$, to be in agreement on $c^\star$, resulting in agreement factors close to 1, thereby reducing uncertainty. In the extreme case where all other patterns are in agreement on $c^\star$ (i.e., $\forall k, A_k = 1$), the summation term equals 1 and $m(\Omega_A) = 0$. In the other extreme case, where all other patterns are in disagreement on $c^\star$ (i.e., $\forall k, A_k = 0$), the summation term equals 0 and $m(\Omega_A) = 1$. These extreme cases allow us to place bounds on the uncertainty due to class agreement.

Additionally, we can define another component of uncertainty based on the disagreement on $c^\star$ between the observed ordinal pattern and the rest of the patterns. First, let D_k be proportion of patterns whose similarity is equal to $k\Delta_s$ and do not have $c^\star$ as their highest probability class. We refer to D_k as the disagreement factor. With this, we define the

uncertainty in the class disagreement between ordinal patterns as

$$m(\Omega_D) = \frac{1}{b} \sum_{k=0}^{K-1} (k+1)\Delta_s D_k. \tag{8}$$

In this equation, we expect dissimilar patterns to be in disagreement on $c^\star$. Therefore, disagreement factors of dissimilar patterns are given a lower weight (i.e., lower importance). In the extreme case where all other patterns disagree on $c^\star$ (i.e., $\forall k, D_k = 1$), $m(\Omega_D) = 1$. At the other extreme, where all other patterns are in agreement on $c^\star$ (i.e., $\forall k, D_k = 0$), $m(\Omega_D) = 0$.

Using Eq. (7) & (8), we compute an overall uncertainty mass as an average of both uncertainty measures,

$$m(\Omega) = \frac{\Omega_A + \Omega_D}{2}. \tag{9}$$

This final uncertainty measure takes into account both class agreement and disagreement. To obtain the masses for the singleton propositions, we normalize the posterior distribution $p(c_j|\pi^\star)$ so that the class masses add up to $1 - m(\Omega)$. With mass assignments of the singleton propositions and Ω, we have mass assignment function $m_r(\cdot) = m(\cdot)$, for all r, that satisfies the properties in Eq. (6).

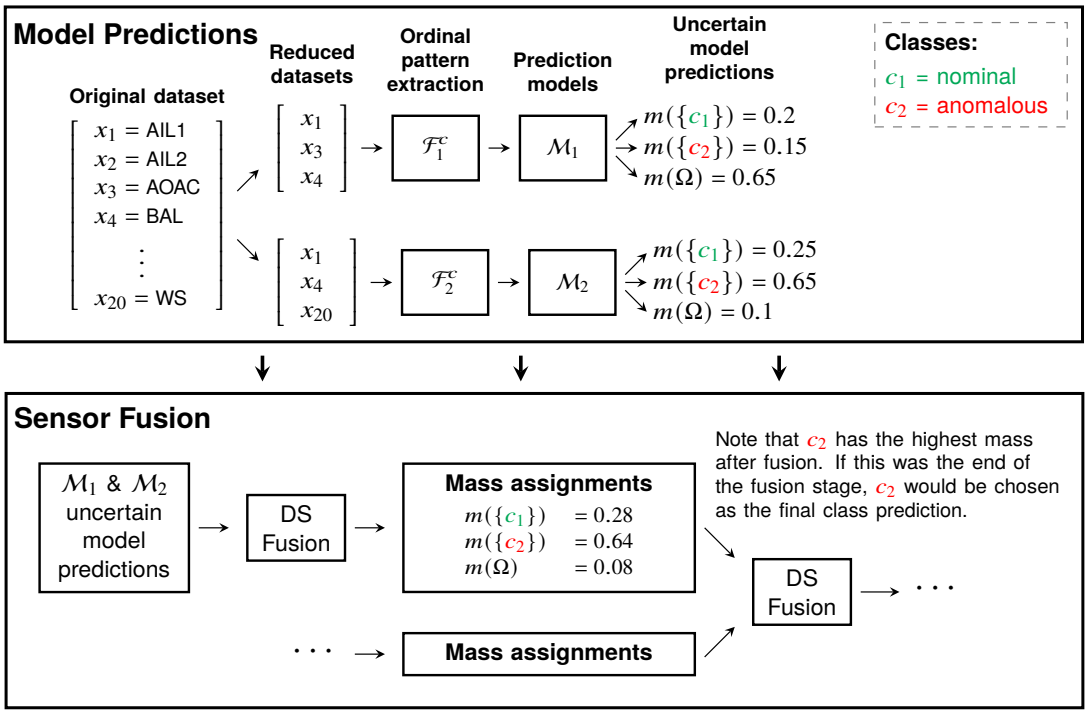


Fig. 6 Illustrative example of the sensor fusion strategy proposed in this study using only two classes. (Top box) From an original dataset containing dozens of sensor readings of the flight dynamics, reduced datasets are created using 3 sensors. The ordinal patterns of these sensor readings, $\mathcal{F}_r^c$, are used to create the ordinal pattern likelihood distributions used in $\mathcal{M}_r$. Uncertainty in a model's class prediction is quantified using DST by assigning a mass to Ω, as seen in the predictions of the models. (Bottom box) Uncertainty in the true class prediction is resolved through the application of multiple stages of DCR. This fusion stage helps the model arrive at a final prediction with reduced uncertainty.

3. Fusion

Information from multiple BoEs can be fused using operators such as Dempster's combination rule (DCR) [20]. In this study, we use the following DCR equation to fuse information from multiple BoE's, a pair at a time:

$$m(A) = \frac{\sum_{B \cap C = A \neq \emptyset} m_1(B)\, m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B)\, m_2(C)}, \tag{10}$$

where $(A, B, C) \subseteq \Omega$ and m_1, m_2 satisfy Eq. (6). DCR is both associative and commutative, allowing one to fuse pairs of BoEs in any order. From the resulting fused FoD, the class with largest mass is chosen as the final predicted class. Figure 6 is provided as an example and high-level overview of our fusion methodology.

III. Results & Discussion

Our results and discussion are broken up by the following research questions (RQs):

RQ1) What is the level of prediction performance increase that we can observe using DS fusion compared to a more traditional majority vote approach?

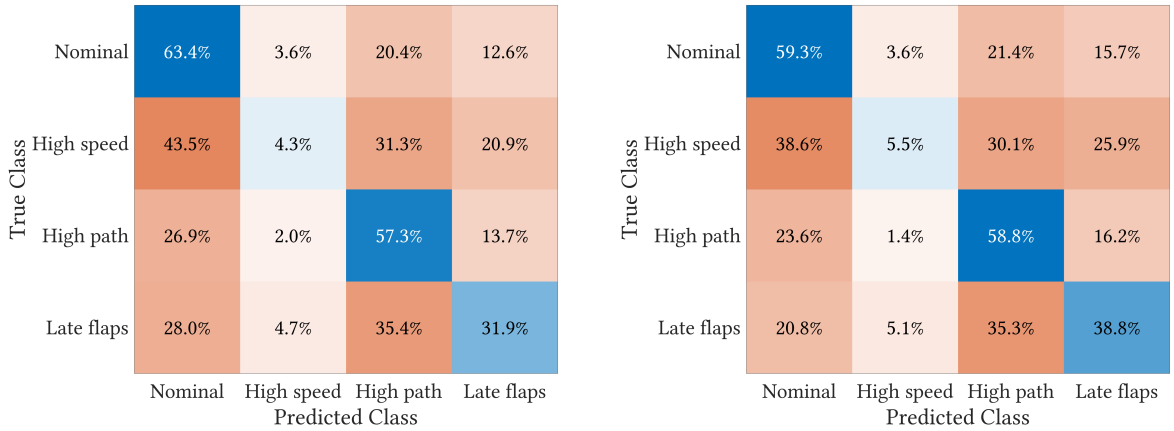
RQ2) How can the variable combinations we generated help us identify the ordinal patterns closely linked to anomalies?

The focus of RQ1 is to analyze the predictive performance increase of DS fusion over the majority vote method. RQ2 discusses the ordinal patterns linked to anomalies to elaborate on their interpretability.

1. RQ1—Performance Increase of DS Fusion Method

The performance of our sensor fusion methodology, simply called Dempster-Shafer fusion (DSF), was analyzed using the following performance metrics: balanced accuracy (BA), F1 score, Matthews Correlation Coefficient (MCC), overall MCC [21], and overall F1 score (micro-averaged) [22]. To establish a baseline performance to compare against, an ensemble majority vote was used to classify without the use of fusion. This method chooses the class with the highest probability from the posterior distribution $p(c_j|\pi^\star)$, where $\pi^\star$ is the observed pattern, as the winning class for each variable combination. The winning classes of all 20 variable combinations are then passed through a majority vote. We refer to the majority vote method as the "no fusion" (NF) method. The performance of NF was also analyzed using the same performance metrics as DSF. We selected four different time instances prior to the 1,000 ft height AGL crossing to analyze the performance of both DSF and NF methods. Their performance is summarized in Figure 7 and Table 2.

The confusion matrices in Figure 7 are true class-normalized, meaning that along each row, the percentages represent the proportion of the true labels that were predicted as one of the four classes. An equal number of test flights were sampled from each class to create the confusion matrices, approximately 190 per class. This was done to avoid the class imbalance from greatly skewing the classification results. Figures 7(a) & (b) show the classification performance at the start of the flight data, 160 seconds prior to landing. Figures 7(c) & (d) show the performance at 100 seconds prior to landing, a few seconds prior to the mean time at which the aircraft crossed the 1,000 ft altitude threshold. From the confusion matrices, we can observe a general improvement in classification performance for both NF and DST as we get closer to 1,000 ft height AGL. The general performance improvement observed in the confusion matrices is supported by an increase in overall MCC and F1 score from 160 to 100 seconds prior to touchdown for both methods. MCC is a performance metric that measures the correlation between the ground truth (i.e., true labels) and predicted labels. MCC ranges from -1 to 1, where 1 represents total agreement between ground truth and predictions and -1 represents total disagreement. A value of 0 for MCC represents a classifier that performs no better than random chance. The overall MCC provided in the confusion matrices and in Table 2 is an extension of the traditional binary-class MCC to the multiclass problem with more than two classes [21]. In particular, we note the larger MCC exhibited by DSF compared to NF at 100 seconds prior to landing. This difference translates to a 15% increase in overall MCC for DSF compared to NF. A percent change improvement of 7.5% in F1 score is also reflected in DSF over NF at 100 seconds prior to landing. Despite the poor classification accuracy of high speed anomalies by DSF, especially at 100 seconds prior to landing when compared to NF, the boost in performance in other classes (such as high path and late flaps) managed to improve the overall MCC and F1 score of DSF versus NF. To shed more light on the misclassifications, we turn our attention to the results in Table 2.



(a) NF at 160 s prior to landing (MCC = 0.20, F1 = 0.39)

(b) DSF at 160 s prior to landing (MCC = 0.22, F1 = 0.41)

(c) NF at 100 s prior to landing (MCC = 0.39, F1 = 0.54)

(d) DSF at 100 s prior to landing (MCC = 0.45, F1 = 0.58)

Fig. 7 Confusion matrices using (a) no fusion (NF) at 160 s prior to landing, (b) DS fusion (DSF) at 160 s prior to landing, (c) NF at 100 s prior to landing, and (d) DSF at 100 seconds prior to landing. These predictions are made before the 1,000 ft height AGL crossing, which at occurs at a mean time of 93 s prior to landing. The overall MCC and F1 score values are provided for each confusion matrix.

Table 2 shows a comparison of all performance metrics at four different time instances during the last 160 seconds of flight. The time instances are spaced out in 20-second intervals, beginning from the first time step in the flight data. The results at 160 seconds and 100 seconds in Table 2 correspond to the confusion matrices in Figure 7. In particular, we want to use the evolution of the performance metrics in the table to attempt to explain 1) why, in both methods, high speed has a poor classification rate compared to the other classes and 2) why some anomalies tend to be erroneously classified as nominal by NF. We need to look at the per-class MCCs and F1 scores to answer the first question. These two metrics show that high-speed predictions have the lowest performance values out of all four classes. The high speed prediction performance is especially poor at 160 seconds prior to landing, where F1 score is ≤ 0.10 and MCC is ≤ 0.05 for both methods. The MCCs and F1 scores for high speed tend to improve as we get closer to landing, with NF having a strong lead of 0.45 for MCC, compared to DSF's 0.36 at 100 seconds prior to landing. We hypothesize that the overall poor performance of both methods at classifying high speed comes down to the lack of distinct enough likelihood distributions for high speed and other classes. This potential explanation is supported by the fact that a large percentage of high-speed flights were incorrectly classified as nominal, as observed in Figure 7. The misclassification of anomalous classes as nominal naturally leads to a discussion of the second question. This observed tendency may be a consequence of the class imbalance, where the nominal class makes up 90% of the data. Despite balancing the test data, the wide

gamut of patterns observed in the "catch-all" nominal class may be negatively impacting performance. Despite the poor performance on high speed, DSF improves the classification accuracy of the two minority classes: high path and late flaps setting. In general, the performance improvement of DSF versus NF tends to increase across most performance metrics as we get closer to landing, as shown in Table 2.

Table 2 Classification performance metrics using no fusion (NF) and DS fusion (DSF) at four time instances prior to landing.

Performance Metric	Class	Time Prior to Landing							
		160 s		140 s		120 s		100 s	
		NF	DSF	NF	DSF	NF	DSF	NF	DSF
BA	Nominal	0.65	0.66	0.69	0.72	0.72	0.75	0.71	0.75
	High speed	0.50	0.51	0.54	0.55	0.54	0.54	0.64	0.60
	High path	0.64	0.65	0.65	0.66	0.69	0.70	0.75	0.78
	Late flaps	0.58	0.60	0.62	0.63	0.67	0.71	0.68	0.76
MCC	Nominal	0.27	0.29	0.33	0.39	0.39	0.44	0.38	0.47
	High speed	0.02	0.05	0.13	0.16	0.12	0.14	0.28	0.29
	High path	0.25	0.27	0.29	0.30	0.37	0.38	0.49	0.52
	Late flaps	0.18	0.20	0.26	0.27	0.35	0.41	0.44	0.50
F1 Score	Nominal	0.48	0.49	0.53	0.56	0.57	0.60	0.56	0.61
	High speed	0.08	0.09	0.19	0.22	0.22	0.22	0.45	0.36
	High path	0.47	0.48	0.48	0.49	0.53	0.55	0.62	0.65
	Late flaps	0.36	0.39	0.42	0.45	0.50	0.56	0.53	0.63
MCC (Overall)	—	0.20	0.22	0.26	0.29	0.32	0.36	0.39	0.45
F1 Score (Overall)	—	0.39	0.41	0.44	0.46	0.48	0.51	0.54	0.58

2. RQ2—Ordinal Patterns Linked to Anomalies

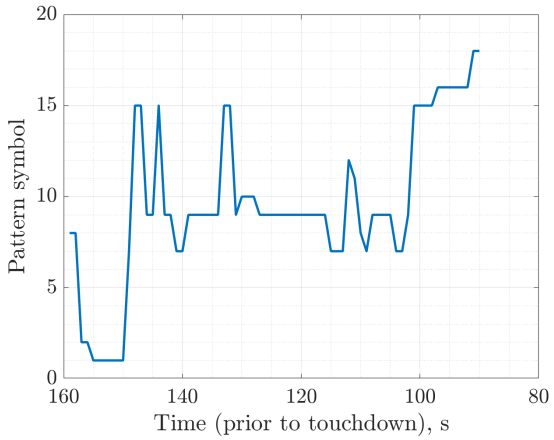
The use of ordinal patterns and multiclass classification based on sensor fusion provides a unique opportunity to identify the ordinal patterns present before the onset of an anomaly. The representation of an aircraft's physical multivariate state by ordinal patterns also allows us to touch on its interpretability. Before extracting the ordinal patterns linked to anomalies, we first need to identify the sensors that may have contributed to the overall good classification performance of DSF. We note that our use of the word sensor refers to a model/BoE $\mathcal{M}_r$ that is used to define initial mass assignments for some variable combination r. To identify variable combinations that produced good sensors, we hypothesized that sensors that made the most correct predictions would possess a lower median uncertainty mass, $m(\Omega)$, for the correct class compared to the other classes. As a consequence of the design of the DSF method, each of the $R = 20$ models can be used individually to classify a flight. This is done by simply taking the class corresponding to the highest singleton mass as the winner. However, in doing so, we ignore the uncertainty component $m(\Omega)$, which captures the degree of "trust" on the singleton masses extracted from the likelihoods $p(\pi|c)$. When adequately fused, the uncertainty component can be used to improve classification performance, as we were able to show in RQ1. To take uncertainty into account, we selected the variable combinations that not only had the most correct class predictions based on the singleton masses, but also had the lowest median uncertainty—calculated across all correctly predicted test flights—at a time step of interest. To test our hypothesis using the variable combination selection criteria, we focused on the predictions of the minority class late flaps at 100 seconds prior to landing. This involved an analysis of all 20 variable combinations in testing fold 3. The variables combinations that make up testing fold 3 are listed in Table 4 in Appendix V.A. To verify the hypothesis quantitatively, we performed a Wilcoxon rank-sum test to test whether the median of the uncertainty mass distribution at 100 seconds before landing for late flaps was significantly different from that of the other classes. For all comparisons between late flaps and the other classes, the rank-sum test yielded

$p < 0.001$, which meant there was a significant difference in the median uncertainty.

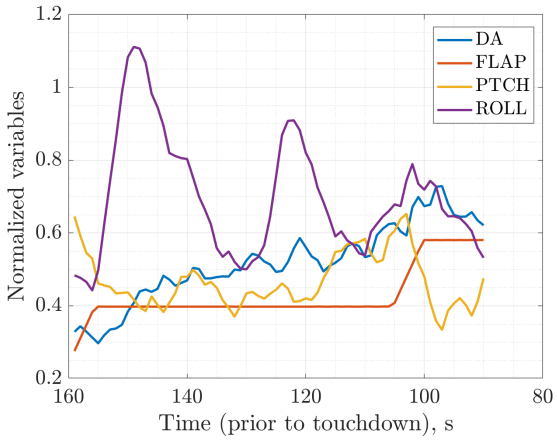
We selected combination 11 after applying the selection criteria to all variable combinations in testing fold 3. To visualize the results, we plotted the evolution of the uncertainty distributions in Figure 8(c) at all time steps in the flight data. At 100 seconds prior to landing, the median uncertainty mass of the BoE is 0.43. For comparison, the median mass uncertainty of the same BoE on classes nominal, high speed, and high path was 0.91, 0.94, and 0.94, respectively. All of this prior analysis was necessary to justify our inspection of the ordinal patterns created by variable combination 11 in fold 3. With the justification out of the way, we plotted the ordinal pattern evolution for a test flight, which was correctly classified by combination 11 in Figure 8(a), along with corresponding normalized variables in Figure 8(b). To understand the figure, each of the 24 possible ordinal patterns was given a unique number. The complete list of assigned numbers or symbols is found in Table 5 in Appendix V.B. In Figures 8(a) & (b), we can observe how each pattern change corresponds to a state transition in the test flight. A change in the pattern models an interaction between flight variables. Thus, the physical dynamics of the aircraft are captured by an ordinal pattern. We also note that throughout the 70-second time window captured by the flight data, different patterns tend to occur at different stages of the approach and landing phase. An analysis of patterns that occur prior to an anomaly (i.e., precursor events) can serve as a method for identifying indicators of an imminent high-risk anomaly. These indicators can then be targeted during pilot training and mission planning to prevent the aircraft from transitioning into a high-risk pattern/state. Again, due to the close link between ordinal patterns and an aircraft's physical state, high-risk indicator patterns possess a higher degree of interpretability compared to the features of more complex models.

IV. Conclusion

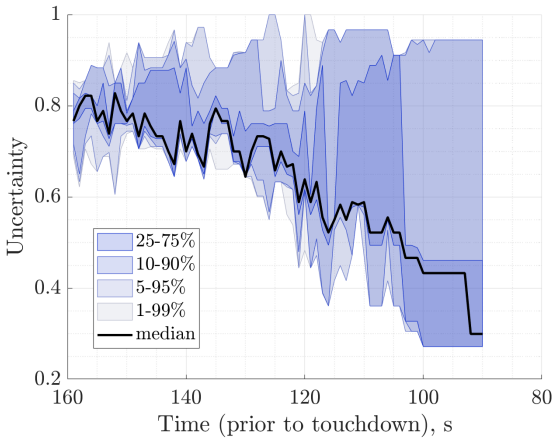
The field of anomaly detection and classification in aeronautical applications is driven by ML and DL models that rely on latent features for accurate classification. In this study, we presented a method for classification that takes advantage of interpretable features, ordinal patterns, to create less complex classifiers with the aid of sensor fusion. By creating a collection of different flight variable combinations, we were able to use DST and a measure of pattern similarity to define an uncertainty measure around the likelihood distributions of each class. Through this uncertainty quantification, our approach takes into account that some variable combinations may be better tailored for the classification of certain anomalies. The results of our fusion method DSF showed an improvement across all performance metrics compared a majority vote approach. Close to the 1,000 ft height AGL crossing, we observed a percent change improvement of 7.5% for overall F1 score and 15% for overall MCC. This classification performance improvement can help reduce the number of incidents and accidents related to unstable approaches during landing. To continue developing the fusion methodology in the future, our goal is to design a mass assignment function capable of quantifying uncertainty in sets of more than one class (doubletons, tripletons, etc.). Moreover, the use of more than 4 variables per variable combination may result in an increased number of ordinal patterns that can be used to more accurately identify the presence or absence of an anomaly. Other potential improvements include an analysis of the performance when training and testing on the full imbalanced dataset, as well as a detailed comparison to other fusion methods used for classification. With these future improvements, we will be able to properly compare our methodology to other state-of-the-art ML/DL anomaly detection methods. This will be done using the same anomalies and datasets to provide a fair comparison. In conclusion, the results presented in this study enable us to identify flight dynamics/states associated with rare high-risk anomalies. Avoiding these states can potentially reduce risk in airframes, other than commercial aircraft, operating with varying levels of human autonomy.



(a) Ordinal pattern evolution



(b) Normalized variables evolution



(c) Uncertainty evolution

Fig. 8 Ordinal pattern results of variable combination 11, fold 3, that exhibited a lower median uncertainty for the late flaps class compared to other classes. (a) Ordinal patterns observed in a test flight that was correctly classified by the BoE of the variable combination. (b) Normalized flight variables corresponding to observed ordinal patterns. (c) Evolution of uncertainty mass, $m(\Omega)$, distribution.

V. Appendix

A. List of Flight Variables in Curated 4-Class Dataset

Table 3 provides the list of all 20 variables found in the 4-class dataset.

Table 4 provides the list of all $R = 20$, 4-variable combinations used to create models/BoEs $\{\mathcal{M}_r\}_{r=1}^{R}$ for each fold.

B. Pattern to Symbol Mapping

The standard pattern to symbol mapping used in this study for all patterns of length 4 is provided in Table 5.

Table 3 Information of all flight variables contained in flight data. The index of the variable is also the symbol used in the ordinal pattern.

Index/Symbol	Variable Description	Identifier/Name	Unit
1	Left aileron position	AIL1	degree
2	Right aileron position	AIL2	degree
3	Angle of attack	AOAC	degree
4	Altitude (baro corrected)	BAL	ft
5	Airspeed	CAS	kn
6	Selected course angle	CRSS	degree
7	Drift angle	DA	degree
8	Left elevator position	ELEV1	degree
9	Flap position	FLAP	discrete count
10	Glideslope deviation	GLS	percentage
11	Selected heading angle	HDGS	degree
12	Localizer deviation	LOC	percentage
13	Core average speed (engine)	N2	percentage
14	Total pressure	PT	mbar
15	Pitch angle	PTCH	degree
16	Roll angle	ROLL	degree
17	Rudder position	RUDD	degree
18	True heading angle	TH	degree
19	Vertical acceleration	VRTG	g
20	Wind speed	WS	kn

Table 4 All 20, 4-variable combinations used to create the BoEs (bodies of evidence) for each fold. Only the index of each variable is provided (see Table 3 for variable description).

Comb.	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1	9, 10, 15, 20	7, 9, 10, 20	3, 7, 12, 20	4, 7, 10, 15	7, 10, 12, 16
2	4, 7, 10, 12	10, 14, 15, 16	7, 10, 12, 14	12, 13, 14, 15	10, 13, 15, 16
3	7, 15, 16, 20	3, 9, 10, 16	7, 10, 16, 20	7, 9, 15, 20	7, 12, 14, 16
4	3, 7, 12, 13	3, 7, 9, 20	7, 9, 12, 15	9, 10, 15, 16	7, 10, 12, 13
5	10, 12, 16, 20	7, 12, 13, 15	3, 9, 10, 20	7, 12, 13, 16	3, 7, 12, 16
6	3, 7, 16, 20	7, 12, 13, 20	7, 9, 16, 20	7, 13, 15, 20	7, 9, 12, 20
7	9, 10, 12, 20	3, 7, 13, 20	5, 7, 10, 16	12, 14, 15, 16	9, 15, 16, 20
8	3, 9, 12, 20	10, 12, 13, 20	12, 15, 16, 20	3, 7, 9, 12	7, 12, 14, 15
9	3, 12, 13, 16	3, 13, 16, 20	10, 12, 13, 15	7, 10, 12, 20	10, 13, 14, 16
10	7, 12, 15, 20	7, 9, 10, 15	10, 12, 14, 15	9, 10, 16, 20	10, 13, 16, 20
11	10, 12, 15, 16	7, 10, 14, 15	7, 9, 15, 16	10, 12, 14, 16	7, 10, 13, 20
12	4, 10, 15, 16	9, 10, 12, 16	3, 10, 13, 20	7, 9, 10, 12	3, 7, 13, 16
13	3, 7, 9, 10	3, 12, 16, 20	4, 7, 10, 16	4, 7, 12, 15	12, 13, 15, 20
14	3, 7, 9, 16	7, 10, 15, 16	9, 10, 12, 15	7, 10, 14, 16	3, 9, 16, 20
15	7, 10, 15, 20	7, 12, 13, 14	13, 15, 16, 20	3, 7, 10, 16	12, 13, 14, 16
16	4, 12, 15, 16	10, 13, 14, 15	10, 12, 15, 20	3, 10, 16, 20	3, 10, 12, 13
17	10, 12, 13, 14	3, 9, 12, 16	7, 13, 14, 15	7, 10, 12, 15	3, 7, 10, 13
18	7, 13, 15, 16	3, 7, 10, 12	4, 10, 12, 16	5, 10, 12, 16	5, 7, 12, 16
19	7, 10, 13, 16	3, 10, 12, 20	7, 10, 13, 14	13, 14, 15, 16	3, 9, 10, 12
20	7, 10, 13, 15	12, 13, 16, 20	3, 7, 10, 20	7, 12, 15, 16	9, 12, 15, 16

Table 5 Standard pattern to symbol mapping for ordinal patterns of length 4.

Pattern	Symbol	Pattern	Symbol
(a, b, c, d)	1	(c, a, b, d)	13
(a, b, d, c)	2	(c, a, d, b)	14
(a, c, b, d)	3	(c, b, a, d)	15
(a, c, d, b)	4	(c, b, d, a)	16
(a, d, b, c)	5	(c, d, a, b)	17
(a, d, c, b)	6	(c, d, b, a)	18
(b, a, c, d)	7	(d, a, b, c)	19
(b, a, d, c)	8	(d, a, c, b)	20
(b, c, a, d)	9	(d, b, a, c)	21
(b, c, d, a)	10	(d, b, c, a)	22
(b, d, a, c)	11	(d, c, a, b)	23
(b, d, c, a)	12	(d, c, b, a)	24

References

[1] NASA, "NASA Aeronautics Strategic Implementation Plan 2023," [Online], 2023. URL https://c3.ndc.nasa.gov/dashlink/projects/85/.

[2] Bleu-Laine, M.-H., Puranik, T. G., Mavris, D. N., and Matthews, B., "Multiclass Multiple-Instance Learning for Predicting Precursors to Aviation Safety Events," *Journal of Aerospace Information Systems*, Vol. 19, No. 1, 2022, pp. 22–36. https://doi.org/10.2514/1.I010971.

[3] Memarzadeh, M., Matthews, B., and Templin, T., "Multi-Class Anomaly Detection in Flight Data Using Semi-Supervised Explainable Deep Learning Model," *AIAA Scitech 2021 Forum*, American Institute of Aeronautics and Astronautics, 2021. https://doi.org/10.2514/6.2021-0774.

[4] Juarez Garcia, E., Mulvihill, M. L., Kharab, M. S., Stephens, C. L., and Napoli, N. J., "Capturing Multivariate Time Series Interactions to Detect High-Risk Instability During Approach," *AIAA AVIATION 2023 Forum*, American Institute of Aeronautics and Astronautics, San Diego, CA, 2023. https://doi.org/10.2514/6.2023-3548.

[5] Oliveira Filho, P. S., "The growing level of aircraft systems complexity and software investigation," [Online], 2018. URL https://www.isasi.org/Documents/library/technical-papers/2018/Wed/The%20Growing%20Level%20of%20Aircraft%20Systems%20Complexity%20and%20Software%20Investigation%20-%20Paulo%20Soares%20Oliveria%20Filho.pdf.

[6] Janakiraman, V. M., "Explaining Aviation Safety Incidents Using Deep Temporal Multiple Instance Learning," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, New York, NY, USA, 2018, pp. 406–415. https://doi.org/10.1145/3219819.3219871.

[7] Napoli, N. J., Stephens, C. L., Kennedy, K. D., Barnes, L. E., Juarez Garcia, E., and Harrivel, A. R., "NAPS Fusion: A framework to overcome experimental data limitations to predict human performance and cognitive task outcomes," *Information Fusion*, Vol. 91, 2023, pp. 15–30. https://doi.org/https://doi.org/10.1016/j.inffus.2022.09.016.

[8] Juarez Garcia, E., Stephens, C. L., and Napoli, N. J., "Detecting High-Risk Anomalies in Aircraft Dynamics Through Entropic Analysis of Time Series Data," *AIAA AVIATION Forum*, Chicago, IL, 2022.

[9] "IATA Unstable Approaches, Risk Mitigation Policies, Procedures and Best Practices," , Oct. 2016.

[10] Federal Aviation Administration (FAA), "Advisory Circular 91-79A – Mitigating the Risks of a Runway Overrun Upon Landing," [Online], 2018. URL https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_91-79A_Chg_2.pdf.

[11] Basora, L., Olive, X., and Dubot, T., "Recent Advances in Anomaly Detection Methods Applied to Aviation," *Aerospace*, Vol. 6, No. 11, 2019, p. 117. https://doi.org/10.3390/aerospace6110117.

[12] Oza, N. C., and Stephens, C., "Data-Driven Safety," *Flight Safety Foundation*, 2021. URL https://flightsafety.org/asw-article/data-driven-safety/.

[13] Komite Nasional Keselamatan Transportasi (KNKT), "Aircraft Accident Investigation Report KNKT.18.10.35.04," [Online], 2018. URL http://knkt.dephub.go.id/knkt/ntsc_aviation/baru/2018%20-%20035%20-%20PK-LQP%20Final%20Report.pdf.

[14] Gaw, N., Yousefi, S., and Gahrooei, M. R., "Multimodal Data Fusion for Systems Improvement: A Review," *IISE Transactions*, Vol. 54, No. 11, 2021, pp. 1098–1116. https://doi.org/10.1080/24725854.2021.1987593.

[15] Shafer, G., *A mathematical theory of evidence*, Vol. 42, Princeton University Press, 1976.

[16] Federal Aviation Administration (FAA), "Safety Briefing: Stabilized Approach and Landing," [Online], 2018. URL https://www.faa.gov/news/safety_briefing/2018/media/se_topic_18-09.pdf.

[17] Vanwinckelen, G., and Blockeel, H., "Look before You Leap: Some Insights into Learner Evaluation with Cross-Validation," *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Workshop on Statistically Sound Data Mining*, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2014, p. 3–19.

[18] Memarzadeh, M., Akbari Asanjan, A., and Matthews, B., "Robust and Explainable Semi-Supervised Deep Learning Model for Anomaly Detection in Aviation," *Aerospace*, Vol. 9, No. 8, 2022, p. 437. https://doi.org/10.3390/aerospace9080437.

[19] NASA, "DASHlink – Sample Flight Data," [Online], 2012. URL https://c3.ndc.nasa.gov/dashlink/projects/85/.

[20] Shafer, G., "Dempster's rule of combination," *International Journal of Approximate Reasoning*, Vol. 79, 2016, pp. 26–40.

[21] Gorodkin, J., "Comparing two K-category assignments by a K-category correlation coefficient," *Computational biology and chemistry*, Vol. 28, No. 5-6, 2004, pp. 367–374.

[22] Narasimhan, H., Pan, W., Kar, P., Protopapas, P., and Ramaswamy, H. G., "Optimizing the multiclass F-measure via biconcave programming," *2016 IEEE 16th international conference on data mining (ICDM)*, IEEE, 2016, pp. 1101–1106.