COMPARATIVE ANALYSIS OF COLLISION AVOIDANCE DECISION-MAKING ACROSS ORGANIZATIONS

Pavithra Ravi[1], Carolin Frueh[2], Petra Chow[3], Miguel Á. Rosique[4], Jan Siminski[5], Okchul Jung[6], Martin Wermuth[7], Saika Aida[7], Ralph Kahle[7], and Hauke Fiedler[7]

[1]*Space Operations and Astronaut Training, DLR, 82234 Weßling , Germany, Email: pavithra.ravi@dlr.de*
[2]*School of Aeronautics and Astronautics, Purdue University, West Lafayette, 47907 IN, USA, Email: cfrueh@purdue.edu*
[3]*HawkEye 360, Inc., Herndon, 20170 VA, USA, Email: petra@pchow.space*
[4]*GMV, 28760 Tres Cantos, Spain, Email: mrosique@gmv.com*
[5]*Space Debris Office, ESA/ESOC, 64293 Darmstadt, Germany, Email: jan.siminski@esa.int*
[6]*SSA Research Office, KARI, 169-84 Gwahak-ro, South Korea, Email: ocjung@kari.re.kr*
[7]*Space Operations and Astronaut Training, DLR, 82234 Weßling , Germany, Email: {martin.wermuth, saika.aida, ralph.kahle, hauke.fiedler}@dlr.de*

ABSTRACT

The growing density of human-made objects in orbit is resulting in an increase in conjunction events. Collision avoidance has become an essential component of satellite operations, with many organizations relying on on-call satellite operators to initiate avoidance maneuvers to mitigate collision risks. Given the expected proliferation of conjunctions in the coming decade, it will be essential to automate this process, thus reducing operator workload and enhancing scalability. This study compares the decision-making processes of satellite operators across five organizations, including both space agencies and private companies. Six analysts are given a set of 30 critical conjunction events and tasked with making *go/no-go* decisions. Their decision criteria and thresholds are analyzed to identify variations in their approaches. Based on these insights, deterministic rule-based classifiers are developed to replicate each analyst's decision-making process. Additionally, a Long Short-Term Memory (LSTM) model, trained on a synthetic dataset of 8000 critical conjunctions, is tested on the same 30 events assessed by the analysts, and its classifications are compared to theirs. The performance of both the rule-based classifiers and the LSTM model is further evaluated on the synthetic dataset, with the LSTM achieving an F1-score of 88%, demonstrating its strong potential for automating or supplementing decision-making processes.

Keywords: Collision Avoidance; Conjunction Assessment; Dempster-Shafer Theory; Rule-Based Classification; Long Short-Term Memory; Artificial Intelligence.

1. INTRODUCTION

As of March 2025, the number of objects cataloged by space surveillance networks reached over 39,000 [1]. 11,100 active satellites operate against this congested backdrop comprising operational payloads and debris alike. Of these, over 8000 satellites operate in Low Earth Orbit (LEO) [2], making this regime a hot spot for critical conjunctions. Given the number of satellites in orbit is expected to exceed 100,000 by 2030 [3], collision avoidance activities will be of utmost importance to minimize the risk of fragmentation events in an already crowded orbital environment. Over 43,000 conjunction events with a probability of collision (PoC) greater than 1E-6 occur monthly in LEO [4], resulting in at least 2 actionable alerts per satellite per week [5]. The frequency of these critical warnings is expected to proliferate, imposing significant pressures on satellite operators who rely on manual, analyst-dependent collision avoidance procedures.

Most of the collision avoidance process is automated – there are models in place for collision screening, calculation of critical parameters at the time of closest approach (TCA), propagation of the states and covariances of the collision pairs to the TCA, and maneuver optimization, if a maneuver is deemed necessary. A key aspect of the process is *go/no-go* decision-making – deciding for or against a maneuver to mitigate the collision risk. While some entities such as SpaceX automate decision-making using PoC thresholds [6], many organizations rely on analysts working on-call around the clock to make decisions when faced with a critical conjunction.

The need for manual *go/no-go* decision-making stems from the high stakes involved – an incorrect decision could, at worst, result in the loss of a satellite. Conversely, excessive maneuvers deplete the fuel budget of the mission and disrupt planned satellite operations. Re-

lying on human analysts to make the *go/no-go* decision allows for more control over the process, particularly in cases that require careful, case-by-case assessment. Although tricky conjunction events currently make up only a small fraction of all cases, they present unique challenges that complicate decision-making. Such conjunction events include those with sudden jumps in PoC or covariance values, cases where values hover around predefined PoC or miss distance thresholds, or situations involving seemingly untrustworthy or error-ridden data. Given the expected rise in conjunctions over the next decade, these difficult cases could still constitute a significant number of events overall – manual decision-making practices are unlikely to feasibly scale to accommodate such a scenario.

In recent years, many works have explored the potential of automating aspects of the collision avoidance process using AI. European Space Agency's (ESA) 2019 Spacecraft Collision Avoidance Challenge tasked participants with predicting the final collision risk (PoC) between two objects using Conjunction Data Messages (CDMs) [7]. AI models for PoC prediction have also been investigated by [8] and [9], where classification models assess events as high or low risk based on a PoC threshold of 1E-6, as used in ESA's competition. The use of deep learning techniques, such as Long Short-Term Memory (LSTM) models, has been explored by [9], [10], and [11] to predict PoC and covariance trends in future CDMs, and for *go/no-go* decision-making [12]. Covariance forecasting using machine learning with diffusion models has also been implemented [13]. Beyond feature prediction, approaches to further automate event filtering and characterization through the use of gradient boosted decision trees, graph neural networks, and genetic algorithms have also been put forth [14, 15, 16] . A multi-class classification framework has been proposed to assess a CDM's likelihood of changing risk category [17], while [18] applies Evidence Theory to categorize encounter geometries into five classes based on event criticality. Additionally, in line with the comparative study presented in this work, prior studies have also explored differences in collision avoidance practices across entities [19, 20, 21].

This work innovates by studying 30 *go/no-go* decisions made by analysts at different organizations and applying the findings towards the development of deterministic and AI models. Two surveys are provided to the analysts – the *preliminary survey* is aimed at obtaining an overview of collision avoidance practices in the company and general perspectives on automating the decision-making process; the *decision survey* notes the decisions made by the analysts, as well as the factors influencing them. Outcomes from the surveys are used to develop rule-based classification models to capture each analyst's decision-making process. These models are then compared to a Long Short-Term Memory (LSTM) model based on their ability to classify events in a simulated dataset of 8000 events, where the ground truth is inherently known. The LSTM model used in this study is a refinement of a previously developed model in [12]. It is trained on 6400 events using an 80-20 train-test split and evaluated on the same 1600-event test set as the rule-based classifiers for a direct comparison. The extent to which *go/no-go* decision-making can be reliably automated using deterministic or AI-based approaches is investigated.

2. DATA SOURCES

2.1. Dataset A – 30 Events

The 30 events evaluated by analysts are based on events from the test set of ESA's 2019 Spacecraft Collision Avoidance Challenge [7]. The events are adapted to be higher risk – the Hard Body Radius is changed to yield a higher Probability of Collision (PoC) and the time to the Time of Closest Approach (TCA) for each CDM is reduced by 0.6 days, to mimic the more critical events faced operationally by the analysts. Details on the complete event generation process can be found in a previous work, [12], for which it was originally developed. Since this project aims to uncover if the decision-making process can be automated, it is of particular interest to assess events which especially need human appraisal – events near PoC/miss distance thresholds and those exhibiting unusual trends. Understanding how analysts evaluate events and make decisions is key to applying these insights in the development of automated decision-making models. Figure 1 shows the PoC distribution in the last CDM across the 30 events. As seen in the figure, these events are high-risk, with PoC values nearing the commonly used 1E-4 threshold.

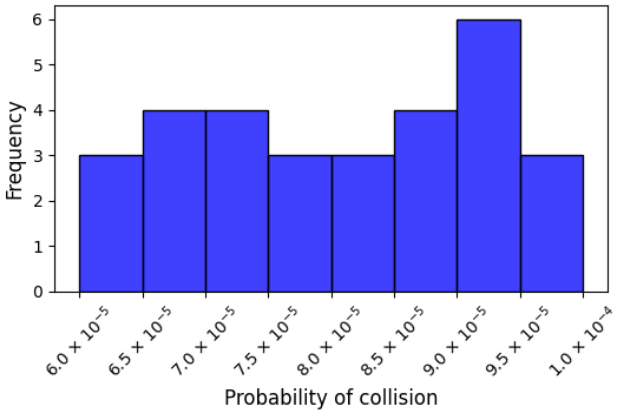

Figure 1: PoC in the last CDM of the 30 events.

The events are provided to the analysts in the Consultative Committee for Space Data Systems (CCSDS) recommended format, which is an industry standard. Additionally, a summary of key quantities of the event, such as time to TCA, miss distance, PoC, and covariance information across the CDMs, is also provided for convenience. This summary is an output of an in-house, Collision Avoidance System (CAS) at the German Space Operations Center (GSOC) [22]. The analysts make their decisions under the following assumptions: the secondary object is non-maneuverable, avoidance maneuvers cannot

be combined with scheduled maneuvers, and the *go/no-go* decision is based on the final CDM received (at least 1.5 days before TCA).

2.2. Dataset B – 8000 Events

Dataset B comprises 8000 simulated critical events in LEO and is used for the training and testing of the LSTM model introduced later in Section 5.2. Unlike Dataset A, where no definitive ground truth exists, Dataset B is explicitly designed by defining the states of conjunction pairs at a TCA. Thus, the final *truth* states of the two objects in a conjunction pair is known and devised to yield a desired number of collisions and misses. Events are designated as collisions if the objects' states lie within a 50-meter miss distance (d) threshold, and are misses otherwise – d is capped at 500 meters for misses to still result in high-risk events. The dataset is balanced with 4000 collisions and 4000 misses.

For the simulated events, CDMs are generated through the following process: the conjunction pairs are back-propagated to an initial time, t_0, 65 hours before TCA. Synthetic radar observations are generated using 25 simulated ground-based sensors positioned around the Earth. These measurements are fed to an Extended Kalman Filter (EKF) for subsequent orbit improvement. An orbit update from the filter is received every 4-10 hours (7 hours on average), and a corresponding CDM is generated. This replicates the cadence at which operators receive CDMs. 3 CDMs are issued per day, for a total of 9 CDMs per event. In this work, only information up to the 5th CDM (about 1.5 days before TCA) is used in the automated models, to mirror the time to TCA values in Dataset A.

The CDMs in this dataset are used by the LSTM and rule-based classification models (Section 5) to make their predictions. As it is known whether the events in this dataset are collisions or misses, this ground truth is used to evaluate the predictions made by the aforementioned models.

Table 1 summarizes the parameters used for generating the synthetic measurements with the EKF. σ refers to standard deviation.

Table 1: Key parameters used for measurement generation and EKF orbit improvement.

Parameter	Value	Unit
Measurement period	120	secs
No. of measurements per period	120	-
Measurement noise, σ	15	m
Process noise, position σ	1E-10	km
Process noise, velocity σ	1E-8	km/s

120 radar observations are collected during each measurement period – one observation is obtained per second. The measurement noise applied is Gaussian with a mean of 0 and a standard deviation of 15 meters. The process noise is kept small as there are no unmodeled dynamics affecting the generated measurements.

Figure 2 depicts the miss distances and PoCs for the collisions and misses in the dataset – these are values from the last CDM issued before a time to TCA of 1.5 days, which is the same time to TCA limit for the events in Dataset A. Many of the misses population have very low PoCs and thus lie outside the frame of the figure. It can be observed that a number of events have miss distances different from the ground truth – the collision events with $d > 50$ m, for instance. These events have not yet converged to the final ground truth and will do so in subsequent EKF updates. The automated models in Section 5.3 are tasked with using this preliminary information as of 1.5 days before TCA, anticipating the continued evolution of the event, and making a *go/no-go* decision accordingly.

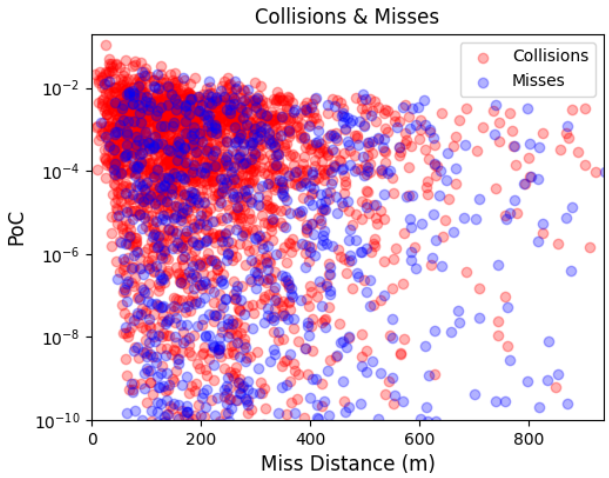

Figure 2: Distribution of events in the generated Dataset B, comprising 8000 events, at TCA - 1.5 days.

3. ANALYST RESPONSES

3.1. Preliminary Survey

Six analysts across five different organizations – the Korean Aerospace Research Institute (KARI), the European Space Agency (ESA), HawkEye 360, Inc., GMV, and the German Aerospace Center (DLR) – participate in this study. A *preliminary survey* has been sent to each organization to gather valuable background information on its collision avoidance practices. Findings from this survey are used to inform the set of events curated for the analysts, as well as the prompts in the *decision survey*, to ensure they are relevant to the processes used at each organization. Key outcomes of this survey are summarized in Table 2. It is observed that decision thresholds

vary, particularly for PoC, with one organization using 1E-3, while the others use 1E-4. All organizations make decisions for satellites in LEO, with some also working in Geostationary (GEO) and Highly Elliptical Orbits (HEO). Notably, the analysts are generally in favor of automating the decision-making process, though some see it as a supplement to existing practices, as opposed to a replacement.

Table 2: Preliminary survey responses for five organizations.

Survey field	Response summary
Number of analysts in the company	2 to 7
Is at least 1 analyst on-call 24/7?	Yes: 3 No: 2
Regimes covered	LEO, GEO, HEO
SSA data sources	18th SDS, LeoLabs, EUSST, Space Data Association
Thresholds used	PoC: 1E-3, 1E-4 d: 100 m
Last CDM considered for decision	12 to 48 hrs before TCA
Would an automated process be beneficial?	*Yes, sooner the better*: 4 *Not yet, but would be valuable in the future*: 1
Is it possible to automate the process?	*Yes, it is deterministic*: 3 *Maybe, it could serve as a filter/second opinion for the analyst*: 2

Additionally, the analysts are asked why the decision-making process could be easy or difficult to automate. The responses recognize both the potential and challenges of automating the decision-making process. While some view automation as feasible, particularly if prioritized, others highlight key obstacles such as the difficulty of quantifying certain factors related to maneuver planning, the uncertainty in available information, and the need to account for unexpected changes. Concerns are raised about ensuring all edge cases are addressed and the operational constraints of coordinating with other spacecraft, which is not currently automated. Despite the possibility of automation, some analysts believe that the severe consequences of potential errors will require human supervision for the foreseeable future.

3.2. Decision Survey

3.2.1. Survey setup

The *decision survey* aims to not only gather the analysts' *go/no-go* responses, but also their reasons for their decisions. Six analysts from the five participating organizations completed the survey.

Table 3 summarizes the information obtained for each event. In addition to the binary *go/no-go* decision, analysts are asked to provide their perception of the event – specifically regarding the event's criticality and the ease of their decision. There is also space for analysts to share any specific factors that contributed to their decision.

Table 3: Components of the decision survey.

Survey field	Options
Time taken	Free response
Impression of the event	1. Very critical 2. Potentially critical 3. Maneuver not needed (but risky) 4. No action needed, uncritical
Go/No-Go decision	Go No-Go
Ease of decision	1. Yes, threshold-based, looks reliable 2. No, event is on the boundary between critical/uncritical 3. No, jumping values and/or inconsistent trends 4. Other (free response)
Specific reason for decision	Free response
CDM parameter importance	1. Very important 2. Somewhat important 3. Unimportant
Usage of all CDMs	Yes No
Additional comments	Free response

To gather the relative importance of CDM parameters, the analysts are also asked to provide an indication of how useful a parameter is in making their decision. These parameters were specifically indicated as being used by the analysts for decision-making in the *preliminary survey*.

A 1 to 3 ranking is used as shown in Table 3. The CDM parameters that are rated are as follows:

- Probability of collision (PoC). There is a field to indicate whether the True PoC or Maximum PoC is

used. The Maximum PoC is typically considered when the K value (covariance scaling factor) is below 1. K $<$ 1 indicates PoC dilution – a reduction in covariance could result in a higher PoC. Thus, some analysts may choose to consider the Maximum PoC when faced with a diluted event.

- Miss distance (d)
- Time to Time of Closest Approach (TCA)
- Relative velocity
- Radial (R) separation. There is a field to indicate if the along-track (T) and/or normal (N) separations are also specifically considered.
- Position uncertainties (R/T/N σ)
- Collision geometry
- Free response. There is a field to highlight any other parameters used and why they influenced the decision.

3.2.2. Survey responses

The responses to the *decision survey* by the six analysts are summarized in this section.

Table 4 shows the number of events classified as *go* and *no-go* by each analyst.

Table 4: *Go*/*no-go* decision breakdown by analyst for Dataset A, consisting of 30 events.

Analyst	No. of *go*	No. of *no-go*
Analyst 1	17	13
Analyst 2	9	21
Analyst 3	14	16
Analyst 4	11	19
Analyst 5	5	25
Analyst 6	1	29

The table highlights a range of decision tendencies. Analyst 6, for instance, appears to be highly conservative, issuing only a single *go* decision across all events, in contrast to Analyst 1, who tends to opt for a maneuver more frequently. The variability across analysts suggests that analyst decision-making tends to be subjective, with each analyst making decisions that are likely in line with individual operational experiences and company protocols. These differences underscore the complexity of automating the decision process, as such a system would have to account for various interpretations of conjunction events based on the end user's unique needs.

Figure 3 visualizes the *go*/*no-go* responses for each event. Additionally, events that are deemed easy and uncritical (ease: 1, impression: 3-4 from Table 3) by more than half of the analysts, are also highlighted.

When considering the highest and lowest decision compatibilities, it is seen that Analysts 5 and 6 reach 87%, followed by Analysts 4 and 5 at 80%; Analysts 3 and 4 agree on only 37% of events, while Analyst 1 aligned with both Analysts 5 and 6 on 40 % of events. This further highlights the discrepancy in decision-making across analysts. 6 out of the 30 events are regarded as either uncritical or easy events by the analysts on average. 5 out of the 6 of these events show a clear majority for the decision, with only Event 23 – regarded 'easy' – having a 50-50 *go*/*no-go* split. This suggests that more alignment between analysts could be found in cases that are uncritical – 2 of the 3 events that yield 100% agreeability across all analysts are both tagged 'uncritical' by the analysts. Easy events – or those for which a threshold-based decision can safely be made – also tend to have more co-alignment in analyst responses; this suggests that the thresholds used by the analysts are similar, which need not always be the case as seen in Table 2.

Section 4 takes a closer look at outcomes from the *decision survey*.

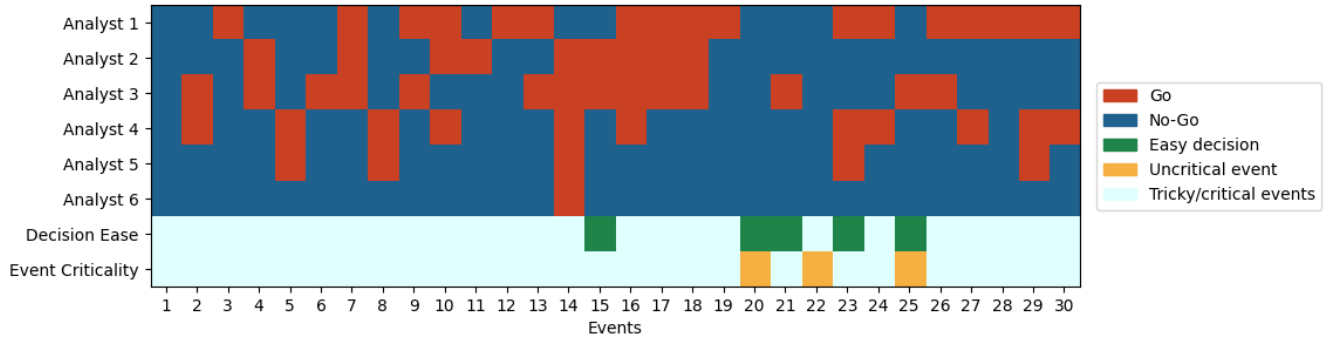

Figure 3: *Go/no-go* responses by analysts for the 30 events of interest.

4. DATA ANALYSIS

4.1. Factors Affecting Decision-Making

Through a qualitative analysis of the *specific reason for decision* in the survey responses, an overview of the key factors dictating the analysts' decisions can be obtained. Table 5 summarizes these rationales. Note, this table does not specifically mention straightforward decisions for nominal cases as these tend to be unanimous – a *Go* decision when thresholds are violated, and everything looks stable, for example.

Unsurprisingly, PoC plays a central role in the decisions with many analysts referring to the PoC threshold and PoC trend behavior in their responses. Additional factors such as miss distance, covariance stability, and PoC dilution are also accounted for. Analysts noticeably differ in their tolerance for uncertainty, with some requiring stable trends before committing to a maneuver, while others maneuver in the face of sudden increases in risk. Certain analysts emphasize operational constraints, such as the required ΔV for maneuver execution, highlighting the practical considerations involved in collision avoidance decisions.

Understanding which features analysts prioritize is essential to select vital features for both the deterministic and AI models developed for decision-making. The 7 *CDM parameter importance* fields in the survey are averaged across the 30 events for each analyst and ranked according to their relative importance in Table 6. Analysts 1 and 6 do not consider collision geometry and relative velocity in their decision-making, and these fields are left blank (—) in the table. Additional features considered by the analysts in the making of their decisions are also listed under 'Free response'.

While Probability of Collision (PoC) is consistently ranked as the most or second-most critical feature by most analysts, there is notable variation in how other features are weighted. Miss distance, time to TCA, and position uncertainties also emerge as important considerations, though their rankings differ across analysts. Some analysts rely heavily on a few key metrics, while others take a more distributed approach, incorporating multiple features into their assessments. Additionally, analysts factor in dilution (through the K value), the maneuver cost, and along-track and normal separations specifically in making their decisions.

It is important to note that a lower ranking does not imply that a feature is unimportant; it simply reflects the relative emphasis placed on it in the overall decision-making process. For example, while relative velocity and collision geometry tend to have lower ranks, they may still play a role in edge cases or when other indicators are inconclusive. The diversity in rankings further highlights the challenge of modeling decision-making through a standardized framework, as analysts weigh different factors based on their personal experience, risk tolerance, and operational constraints.

4.2. Dempster-Shafer Theory for Decision Combination

As observed in Section 3.2.2, analysts often arrive at different conclusions when presented with the same data. If decision-making is to be automated, an important consideration is which human decisions the model should learn from. In the presence of conflicting assessments, a systematic approach to decision fusion is required.

Dempster-Shafer Theory (DST) is used to combine information from different sources and put together a singular outcome based on the supporting evidence for it [23]. In this work, it is used to combine the different decisions made by the analysts to arrive at a single 'superimposed' decision for each event. This approach can be used to establish a ground truth for training AI models or serve as a framework for integrating outputs from multiple in-house models or analysts, ensuring a more robust and unbiased decision-making process.

In the context of this work, several key components of Dempster-Shafer Theory (DST) are defined. The Frame of Discernment (Θ) represents the possible decisions an analyst could make, such as a *go* or *no-go*. In this study, Θ consists of two possible hypotheses: *G* for a *go* decision, and *NG* for a *no-go* decision. The power set in-

Table 5: Decision criteria for *go*/*no-go* maneuvers by analyst.

Analyst	*Go* decision factors	*No-go* decision factors
Analyst 1	PoC suddenly increases (even if below threshold), PoC rises constantly, PoC close to threshold but unstable PoC above threshold with inconsistent PoC/covariance trends	Time to TCA is large, PoC consistently below threshold with no upward trend, PoC close to threshold but unreliable trend, K (covariance scale factor) rising
Analyst 2	Concerning PoC trend, Miss distance near/lower than 100 m	Stable PoC, miss distance, and covariance trends, High R separation despite high PoC, Inconsistent miss distance trends, High covariance
Analyst 3	High PoC with low R separation and unstable T uncertainty, R uncertainty exceeds R separation and PoC is high	Stable RTN separations despite violated thresholds, Unreliable uncertainties
Analyst 4	Low, stable R separation, Potential for PoC increase with improved covariance, PoC near threshold with large uncertainty in T	Stable covariance evolution, Rare but stable event geometry
Analyst 5	Potential for PoC increase with improved covariance, High PoC with low separation (especially in R), without large uncertainties	PoC stable under threshold, PoC highly diluted (results in high maneuver cost due to large uncertainties)
Analyst 6	PoC $>$ 1E-3, Miss distance $<$ 100 m	All other cases

Table 6: Relative feature importance rankings by analyst. 'Free response' indicates additional features that are considered by the analyst. Col. geometry and rel. velocity refer to collision geometry and relative velocity respectively.

Rank	Analyst 1	Analyst 2	Analyst 3	Analyst 4	Analyst 5	Analyst 6
1	PoC	Miss distance	Miss distance	PoC Time to TCA	PoC Time to TCA R separation R/T/N σ	PoC, Miss distance, Time to TCA, R/T/N σ
2	R/T/N σ	PoC	PoC	-	-	-
3	R separation	R/T/N σ	R separation	R/T/N σ	-	-
4	Time to TCA	Time to TCA	R/T/N σ	Miss distance	-	-
5	Miss distance	R separation	Time to TCA	R separation	Miss distance Col. geometry	R separation
6	—	Col. geometry	Col. geometry	Col. geometry	-	—
7	—	Rel. velocity	Rel. velocity	Rel. velocity	Rel. velocity	—
Free response	K value	—	T/N separation	K value	Required dV	—

cludes all the possible combinations of these hypotheses. For instance, it contains elements such as $\emptyset$, which refers to a situation where the event is neither a *go* nor a *no-go*, and $\{G, NG\}$, which represents uncertainty when there is insufficient evidence to decisively choose between *G* or *NG*.

A Basic Probability Assignment (BPA) is then used to quantify the degree of belief assigned to each hypothesis, with the condition that the total mass across all elements of the power set equals 1. The Dempster rule of combination is employed to merge BPAs from multiple sources of evidence, accounting for any conflict or uncertainty between them. By combining BPAs from multiple analysts, this rule generates a single, combined BPA, with the degree of conflict represented by the parameter *K*.

$$m_{1,2}(\emptyset) = 0 \tag{1}$$

$$m_{1,2}(H) = \frac{\sum_{H_i \cap H_j = H} m_1(H_i) m_2(H_j)}{(1 - K)} \tag{2}$$

when $H \neq \emptyset$ and where,

$$K = \sum_{H_i \cap H_j = \emptyset} m_1(H_i) m_2(H_j) \tag{3}$$

Here, m_1 and m_2 are BPAs from two analysts, which are then combined to arrive at a single BPA, $m_{1,2}$ for each combination of hypotheses (H) in the power set. *K* quantifies the degree of conflict between the sources of evidence – the analysts, in this case.

Each analyst's decision is mapped to a BPA using information obtained from the *decision survey* – specifically the decision made, ease of decision, and impression of the event, as defined in Table 3.

Table 7 outlines how the BPAs for each analyst are assigned based on decision, ease, and impression combinations.

None of the BPAs are assigned a value of 1 – this ensures that no source (analyst) has absolute evidence towards a hypothesis. If the BPA of a source is 1 for *m(G)*, for instance, the combined *m(G)* following the rule of combination would also be a 1 regardless of the BPAs of the other sources. A single analyst could dominate the decision fusion process in this case. Additionally, the distribution of some mass to *m(G,NG)* allows for the acknowledgment of the intrinsic uncertainty in the decision-making process.

When analysts classify an event as a *go*, while citing the event as both critical and easy (threshold-based decision), the BPA for *m(G)* is maximized at 0.95, with a small mass (0.05) allocated to uncertainty, *m(G,NG)*. For trickier events (ease: 2-4), the uncertainty correspondingly goes up, resulting in a higher BPA for *m(G,NG)*. Events that are assigned *go* by the analysts, which are rated as 'potentially critical', have an even higher uncertainty – these typically refer to events which analysts predict will increase in criticality in following CDM updates. The BPA assignment for *no-go* events is done similarly, such that *m(NG)* is high for uncritical events (impressions: 3-4) and potential critical events which are easy to make decisions for. For trickier events (ease: 2-4), some of the mass for *m(NG)* is distributed to *m(G,NG)* to represent the higher uncertainty in the decision-making process.

Figures 4 and 5 visualize the *m(G)* assigned to each event for each analyst, with the last row in each figure showing the aggregated *m(G)* following the rule of combination. Two scenarios are considered: Firstly, the case including all analysts (Figure 4) and secondly, the case excluding Analyst 6 (Figure 5), whose use of a different PoC threshold heavily affects the the combined results. The fused result when all six analysts' decisions are used, assigns a *go* hypothesis with $m(G) > 0.5$ to two events (14, 16) and an uncertain hypothesis with $m(G) = m(NG)$ to two events (10, 17). When Analyst 6 is not factored in, the number of fused *go* predictions rises to six.

Through the Dempster-Shafer Theory, particularly challenging events to classify can readily be identified – events with combined BPAs split evenly between hypotheses, or those with values near 0.5, indicate more uncertainty in the decision-making process and can be separately investigated.

Table 7: Basic Probability Assignments from decision survey outcomes.

Decision	Impression	Ease	$m(G)$	$m(NG)$	$m(G, NG)$
Go	1	1	0.95	0	0.05
Go	1	2/3/4	0.75	0	0.25
Go	2	all	0.6	0	0.4
No-go	2	1	0	0.9	0.1
No-go	2	2/3/4	0	0.75	0.25
No-go	3/4	all	0	0.95	0.05

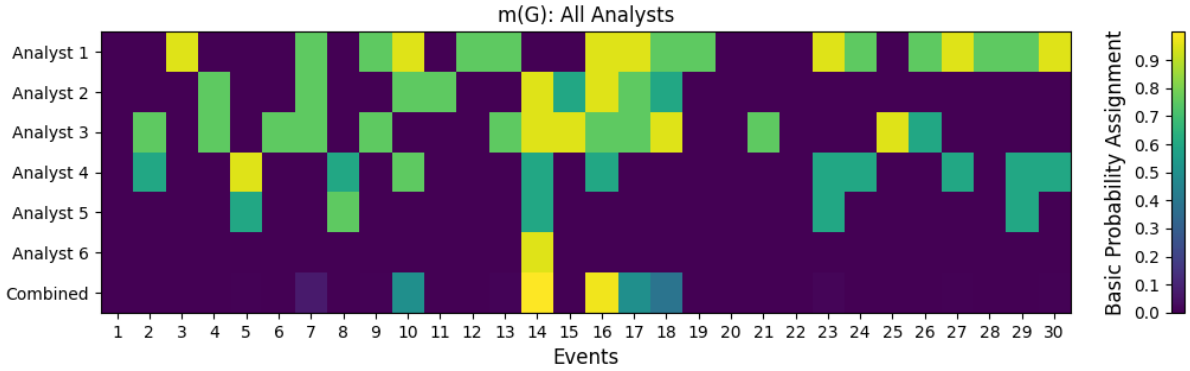

Figure 4: Dempster rule of combination results including all analysts. The combined decision consists of two *go*, 26 *no-go*, and two uncertain cases.

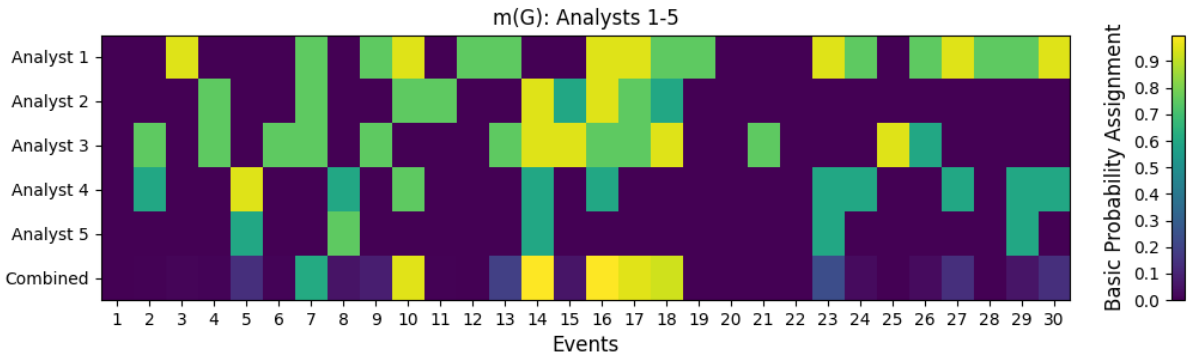

Figure 5: Dempster rule of combination results excluding Analyst 6. The combined decision consists of six *go* and 24 *no-go* cases.

5. DECISION PREDICTION

5.1. Rule-Based Classification

It is of interest to extract insights provided by the analysts in the *decision survey* and transform them into quantitative models that can aid the decision-making process. A rule-based classification approach is proposed, where a set of rules is developed based on each analyst's classifications to produce the same *go* or *no-go* decisions made by the analyst. Six such distinct deterministic models are developed – one for each analyst. To aid generalization of the rule-based classifiers, most rules are selected only if they apply to more than one event. However, this approach is not feasible for analysts with few positive (*go*) decisions, such as Analysts 5 and 6 who assigned 5 and 1 *go*'s respectively.

Table 8 summarizes the decision rules used in the rule-based classifiers, which are derived from the survey responses of the analysts. Note that R, T, N refer to the radial, along-track, and normal components, in the relative position frame centered on the primary object (the object typically monitored by the satellite operator). The covariance terms (with σ) pertain to the secondary object to aid generalization across missions. Unless *last 3* or other temporal condition is specified, the features in the rules pertain to those in the last issued CDM. Mention of reliability in the rules corresponds to a check to ensure that there are no significant jumps in the feature in the

last three CDMs.

PoC, time to TCA, and miss distance, are notably the only features that can serve as standalone decision criteria – high PoC values in the last few CDMs, low miss distance, or a sufficiently large time to TCA alone can be enough to dictate a *go* or *no-go* decision. In contrast, other features, such as RTN separation and covariance, require a combination of factors – always involving at least PoC – to trigger a *go* decision. Each model achieves a 100 percent accuracy rate for the analyst it is calibrated against as it is essentially overfitted to the analyst's decisions.

The evident diversity in decision criteria used by analysts highlights the complexity of automating decision-making. Rule-based classifiers, which can be structured according to individual analysts' preferences, offer a way to automate decision-making while still adhering to the analysts' decision logic and retaining their influence. However, given the significant variability in decision rules, a one-size-fits-all rule-based classifier is unlikely to be effective. Instead, multiple classifiers may need to be developed and tailored for different scenarios or end users with varying risk tolerances and operational priorities. It is worth noting that rule-based classifiers can also be time consuming to develop. Edge cases may be difficult to identify, making such classifiers ineffective for events not explicitly accounted for.

5.2. AI Model

AI models are widely used for classification applications and could potentially be a more effective alternative to rule-based classifiers. Long Short-Term Memory (LSTM) networks, are particularly well-suited for the *go/no-go* classification task due to their ability to capture temporal dependencies and sequential patterns in data. A series of Conjunction Data Messages (CDMs) is available for each event, providing valuable information on the evolution of PoC, miss distance, and other key parameters. LSTMs can recognize trends and fluctuations in these parameters over time due to their complex internal gating mechanisms [24].

A prior study by the author demonstrates the potential of Long Short-Term Memory (LSTM) models for *go/no-go* decision-making on a smaller dataset of 200 events, labeled by analysts at DLR [12]. Deep learning models such as LSTM typically require larger amounts of training data than traditional machine learning models as they have a higher number of trainable parameters and more complex internal memory mechanisms. Labeling thousands of events would be tedious and time-consuming work for analysts. As a result, a synthetic dataset consisting of 8000 events (Section 2.2), which does not need to be labeled by analysts, is developed and used for this work.

5.2.1. Feature engineering and hyperparameter optimization

Survey outcomes (Section 3.2) and findings from the development of the rule-based classifiers (Section 5.1) are used to inform feature-selection for the LSTM model. By selecting features most relevant to analysts' decisions and disregarding less important ones, the model can more easily focus on key patterns, without being burdened by additional complexity and noise. This results in the ten features as shown in Table 9. The features cover a range of parameters, including collision probabilities, separation distances, and uncertainty measures.

In Section 5.1 it is identified that the rule-based classifiers align with analyst decisions when using data simply contained in the last three CDMs (the last CDM is still at least 1.5 days from TCA). As a result, the LSTM model is also only shown three CDMs, with the last CDM also being about 1.5 days from TCA. Using three CDMs strikes a balance, enabling the model to still learn from the recent trend, while minimizing complexity from the accumulation of parameters in a lengthy series.

As is done in [12], the features are clipped and scaled to account for outliers, and to enhance learning. Clipping refers to the process of setting an upper and/or lower limit for a feature, and assigning this number to all values beyond this limit. In this work, all features are assigned an upper limit of 3 standard deviations from the mean, except for TCA, which contains no outliers. Additionally, PoC and Max. PoC values are clipped to a minimum value of 1E-6 and log-transformed, to compress the range of values, preventing low-risk CDMs with near-zero probabilities from disproportionately influencing the developed model. Z-normalization is then applied to the clipped and partially log-transformed data, such that it has a mean of 0 and a standard deviation of 1 as is found to be effective in [12].

Hyperparameter optimization is done using a grid search, where the model's performance is comprehensively assessed on all combinations of hyperparameters of interest. A stratified K-fold cross-validation with 5 folds is implemented alongside the grid search, mitigating any potential bias from data partitioning. Each combination of hyperparameters is thus evaluated five times on the full dataset of 8000 events, where a different 1/5 (20 %) is in the test set. It is ensured that the test split contains an equal percentage-split of the two classes. The F1-score [25], which accounts for both the precision and recall performance of the model, is used to identify the best combination of hyperparameters. Precision, recall, F1-score, and other key metrics are defined in Section 5.2.2.

Table 10 summarizes the hyperparameters selected for the LSTM model following Grid Search-based hyperparameter optimization with stratified 5-fold cross validation.

The developed model comprises three layers. An LSTM layer containing 100 neurons processes the sequential in-

Table 8: Decision rules used for rule-based classifiers, grouped by analyst.

Analyst	Rule	Features
Analyst 1	Last 3 PoC $<$ 9E-5: *no-go*	PoC
	PoC is not strictly increasing *and* PoC $<$ 1E-4 *and* K $<$ 1: *no-go*	PoC trend, K value
	PoC $<$ 9E-5 *and* K not significantly lower: *no-go*	PoC, K value trend
	$t >$ 1.7 days: *no-go*	Time to TCA
Analyst 2	$d <$ 150 m: *go*	d
	PoC is significantly higher *and* PoC $>$ 8E-5 *and* $d <$ 500 m: *go*	PoC trend, d
	$d <$ 1 km *and* any of last 3 PoC $>$ 1E-4: *go*	d, PoC
Analyst 3	Any of last 3 PoC $>$ 1E-4 *and* PoC $>$ 6E-5 *and* $d <$ 500 m: *go*	PoC, d
	Mean of last 2 $d <$ 300 m *and* R sep. $<$ 150 m *and* PoC $>$ 6E-5: *go*	d, R sep., PoC
	$d <$ 500 m *and* R sep. $<$ 100 m *and* PoC $>$ 6E-5: *go*	d, R sep., PoC
	R σ and PoC significantly higher *and* $d <$ 500 m *and* PoC $>$ 6E-5: *go*	R σ trend, PoC trend, d
	N sep. $<$ 1.5 km *and* reliable *and* R sep. $<$ 60 m *and* PoC $>$ 9E-5: *go*	N sep. trend, R sep., PoC
Analyst 4	Last 3 PoC $>$ 8E-5: *go*	PoC
	Any of last 3 PoC $>$ 1E-4 *and* PoC $>$ 6E-5 *and* $d <$ 500 m: *go*	PoC, d
	R/T/N sep. $<$ 5 m *and* PoC $>$ 6E-5 *and* $d <$ 1.2 km: *go*	RTN sep., PoC, d
	T $\sigma >$ 10 km *and* PoC $>$ 6E-5: *go*	T σ, PoC
Analyst 5	Any of last 3 PoC $>$ 1E-3: *go*	PoC
	R sep. $<$ 5 m *and* PoC $>$ 9E-5 *and* any of last 3 PoC $>$ 1E-4: *go*	R sep., PoC
	T $\sigma >$ 18 km *and* any of last 3 PoC $>$ 1E-4: *go*	T σ, PoC
	R $\sigma <$ 100 m *and* reliable *and* T $\sigma <$ 10 km *and* last 3 PoC $>$ 8E-5: *go*	R σ trend, T σ, PoC
Analyst 6	Any of last 3 PoC $>$ 1E-3: *go*	PoC

Table 9: Features used in the LSTM model.

Feature	Unit
Probability of Collision	-
Maximum Probability of Collision	-
K value (covariance scaling factor)	-
Time to Closest Approach	days
Miss Distance	km
Radial, Along-track, and Normal Separation	km
Obj. 2 Radial and Along-track σ	km

Table 10: Optimized LSTM hyperparameters.

Hyperparameter	Optimal value
Neurons	100
Dropout	0.2
Learning rate	0.001
Batch size	64

put data. A dropout rate of 0.2 is applied to minimize overfitting – this regularization technique sets 20% of the neurons to a value of 0 during training, preventing the model from overly relying on specific information in the training set to aid generalization. A fully connected final layer consisting of a single neuron with sigmoid activation is used to classify the event. The learning rate strikes a balance between convergence speed and stability – a higher rate could lead to faster convergence, but risks overshooting the optimal solution. The batch size refers to the amount of events the model processes at a time before updating its weights. A learning rate of 0.001 and a batch size of 64 serve as good trade-offs between computationally efficiency and learning stability in this study.

5.2.2. Model performance and learned feature importance

The results from using the optimized hyperparameters are summarized in Table 11. These are averaged over five shuffles of the dataset, where a stratified 5-fold cross validation is applied on each shuffle.

The metrics used in Table 11 to assess the model's performance are accuracy, precision, recall, and the F1 and F2 scores. Accuracy indicates how many predictions are made correctly, providing a general sense of model performance. Given the balanced dataset consisting of an equal number of 1's (collisions) and 0's (misses) in this case, the accuracy is not biased by class imbalance. Precision represents the fraction of predicted collisions that

Table 11: LSTM performance using optimized hyperparameters. The results shown are for training and test sets of 6400 and 1600 events respectively.

Metric	Training set	Test set
Accuracy	0.88	0.88
Precision	0.84	0.84
Recall	0.95	0.93
F1	0.89	0.88
F2	0.93	0.91

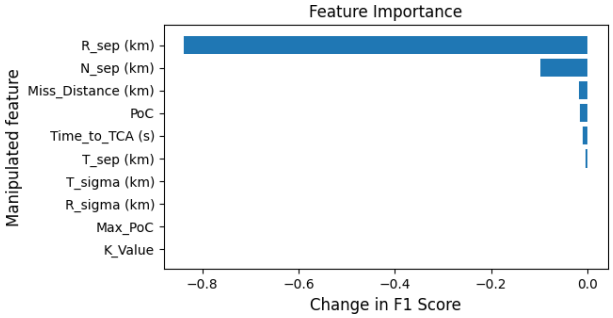

Figure 6: Features ranked according to their contributions to model performance (most significant on top).

are actual collisions. A high precision value suggests the model does not issue a lot of false alarms. Recall captures the proportion of all the collisions identified by the model, reflecting its ability to correctly detect critical events. The F1-score is the harmonic mean of both precision and recall, making it a useful metric when it is of importance to minimize both false positives (true misses incorrectly labeled as collisions) and false negatives (true collisions incorrectly labeled as misses). The F2-score places greater emphasis on recall, prioritizing the identification of all collisions over mitigating false alarms.

The optimized model achieves a high accuracy of 0.88 on both the training and test sets, indicating strong generalization and negligible overfitting. Precision (0.84) and recall (0.93) also score highly on the test set, suggesting that the model effectively identifies most collision events while minimizing false alarms. The test F1-score of 0.88 reflects a good trade-off between precision and recall. The F2-score is particularly relevant in this work, as it places more emphasis on recall – in the context of collision avoidance, false negatives would have far more severe consequences than false positives (potential loss of satellite vs. unnecessary maneuver). The high F2-score of 0.91 in the test set indicates that the model is well-suited for prioritizing safety while maintaining overall reliability.

To acquire more transparency from the model outcomes, it can be beneficial to identify which features affect the model's performance the most. This can be done by running the trained model on a test set where each feature column is manipulated with dummy values, one at a time. This effectively prevents the model from using the true values of the specified feature, such that it is unable to use insights from this feature when classifying the test set. Since the data presented to the model is Z-normalized, the values of the feature to be manipulated are set to 0 – this ensures the data retains its overall statistical properties and does not introduce any new outliers.

Figure 6 demonstrates the feature importance of the LSTM model by summarizing the change in the F1-score when each feature is set to 0. R_sep, T_sep, N_sep refer to the separations in RTN, while R_sigma and T_sigma refer to the covariance terms in the R and T components.

The F1-score with all features available to the model is 0.88. The results indicate that the model heavily relies on distance-based features such as radial separation and normal separation to make its predictions. However, PoC, while favored heavily by analysts in decision-making (Table 6), is found to only have a minimal impact on the LSTM's predictions. The model is able to make almost the same quality of predictions without studying the PoC trend. This model behavior is also observed in [12], where PoC ranked 14 out of 16 features for relative importance. This is likely to occur if there are many events for which the PoC trend of both *go* and *no-go* events is fairly similar, forcing the model to seek out distinctive patterns amongst other features. These events can be identified in Figure 2, where both high PoC misses and low PoC collisions are visible. The LSTM model is trained and tested on a dataset where *go*/*no-go* labels are derived from miss distance (Section 2.2). It is particularly promising that the model places significant importance on distance-based features, effectively focusing on the key factors that influence the ground truth, while disregarding other potentially correlated features that do not directly affect the decision in this setup.

5.3. Comparison of Models for Decision-Making

5.3.1. LSTM vs. analysts' survey responses

The updated LSTM model is tested on the 30 events classified by the analysts to assess its comparative performance. As these events are not part of the synthetic dataset, there is no ground truth against which the predictions can be compared. Hence, the LSTM's performance is evaluated by comparing its predictions against each of the analysts' decisions to observe the extent of alignment as shown in Table 12. The last row in the table also includes a comparison to the combined DST results using Analysts 1-5 (Figure 5).

Since each analyst assigned a different number of *go* and *no-go* decisions, accuracy is not a reliable metric for evaluating decision agreement here. Instead, the F1-score provides a more reliable measure of how well the LSTM model aligns with each analyst's decisions. The LSTM

Table 12: LSTM predictions evaluated against analyst predictions for 30 events. The performance metrics used are accuracy (Acc.), precision (P), recall (R), and the F1 and F2 scores.

LSTM vs.	Acc.	P	R	F1	F2
Analyst 1	0.50	0.63	0.29	0.40	0.33
Analyst 2	0.70	0.50	0.44	0.47	0.45
Analyst 3	0.53	0.50	0.29	0.36	0.31
Analyst 4	0.70	0.63	0.45	0.53	0.48
Analyst 5	0.70	0.25	0.40	0.31	0.36
Analyst 6	0.77	0.13	1.00	0.22	0.42
Combined	0.73	0.38	0.50	0.43	0.47

Table 13: Comparison of model performance on a test set of 1600 events. The performance metrics used are accuracy (Acc.), precision (P), recall (R), and the F1 and F2 scores.

Model	Acc.	P	R	F1	F2
LSTM	0.88	0.84	0.93	0.88	0.91
RBC 1	0.65	0.80	0.40	0.53	0.44
RBC 2	0.69	0.79	0.51	0.62	0.55
RBC 3	0.60	0.79	0.28	0.41	0.32
RBC 4	0.63	0.82	0.33	0.47	0.37
RBC 5	0.66	0.87	0.38	0.53	0.43
RBC 6	0.64	0.84	0.33	0.48	0.38

model aligns most closely with Analyst 4, achieving the highest F1-score of 0.53 – this still signals substantial divergence in decision-making. The model does achieve a high recall score of 1.0 with Analyst 6, indicating the model correctly identifies all *go* decisions assigned by this analyst. However, given that Analyst 6 only assigned one *go*, this high recall is trivial and does not indicate strong agreement in a broader sense. The overall trend suggests that the LSTM model is making very different decisions compared to the analysts. This is likely due to the model being trained on a different dataset (Section 2.2) and relying on its own learned decision criteria. In contrast, the analysts' decisions are based on their individual thresholds for PoC, miss distance, and uncertainty tolerance. Unless the LSTM model is specifically trained on decisions made using these criteria, it is unlikely to align with the analysts' responses.

5.3.2. LSTM vs. rule-based classification

To examine the broader applicability of deterministic models, it is of interest to evaluate their ability to classify unseen events, particularly in comparison to AI models. To explore this, the six rule-based classifiers developed in Section 5.1 are tested on the events from the synthetic dataset generated in Section 2.2. For a fair comparison with the LSTM model, all models are evaluated on the same test set, comprising 1600 events. Table 13 presents a comparison of how each model classifies the synthetic events. Note that RBC refers to 'rule-based classifier', with the number after it corresponding to the analyst it is calibrated to.

The LSTM model generally outperforms the RBCs, achieving an accuracy of 0.88, precision of 0.84, recall of 0.93, and F1 and F2 scores of 0.88 and 0.91, respectively. In contrast, the performances of the RBCs vary significantly. RBC 2, which is calibrated to Analyst 2, performs relatively well with the highest F1-score across all the RBCs with 0.62. Analyst 2 heavily favors miss distance in decision-making, d, (Table 6) and the rules developed for RBC 2 place heavy emphasis on d as a result. Given the target labels for the dataset are also set based on the true d at TCA between the two objects, it is is not entirely surprising that a distance-based classifier performs better on the dataset. RBC 3, RBC 4, and RBC 6 have lower accuracy and F1 scores, reflecting their relatively limited ability to generalize across the synthetic dataset. All three of the RBCs appear to result in many false negatives, resulting in noticeably lower recall. However, all the RBCs tend to have decently high precision values ranging from 0.79 to 0.87. RBC 5 achieves the highest precision (0.87) in the comparison, but this comes at a cost of low recall. The model is selective in making *go* decisions resulting in fewer false alarms, but fails to detect many actual collisions. In comparison, the LSTM model appears well-balanced – it misses few collisions, while mitigating the number of false alarms. Given the models are making predictions based on CDM information capped at TCA - 1.5 days, it can be inferred that the LSTM model is better at anticipating the risk evolution of the events. True collisions that initially appear to be less risky are identified correctly far more often by the LSTM model than by the RBCs.

It must be noted that the RBCs are 'trained' on the (different) dataset of only 30 events classified by the analysts, while the LSTM model is trained on significantly more events (6400) in the same events dataset. While RBCs informed by thousands of analyst decisions would likely perform better, challenges remain in collecting such a large volume of decisions and effectively fitting rules to capture trends across the events without becoming overly complex. Additionally, RBCs are time-consuming to develop as they need to be individually designed and calibrated to an analyst's decision criteria. LSTMs on the other hand, are able to identify and learn patterns in the data on their own, leading to high performance metrics without the need for labor-intensive rule-creation.

The RBC and LSTM decisions (Table 13) generally show greater alignment compared to the decisions made by analysts and the LSTM model (Table 12). This could be due to the small sample size of 30 events in the comparison with the analysts which could result in easily biased

metrics. Another important factor is the inherent variability in human decision-making. Analysts may not always make the same decision when faced with the same event, especially in borderline cases where one might lean on intuition. Humans are susceptible to inconsistencies and errors in judgment, introducing a level of subjectivity that is difficult to account for. In contrast, RBCs follow predefined, consistent rules, making them less prone to such variability. As a result, RBCs may be quite useful as a pre-filter for decision-making, where events unlikely to evolve into critical events can be identified and removed from consideration.

Overall, LSTM networks appear to be a promising solution for decision-making in this context. Their ability to learn directly from data enables them to capture complex decision patterns without requiring manually defined rules. Additionally, their scalability and adaptability make them well-suited for handling larger datasets and evolving decision criteria, reducing the reliance on rigid, predefined rules that may not generalize well across events.

6. CONCLUSIONS

Automating the collision avoidance decision-making process is an essential, but complex task, as it requires careful assessment of multiple parameters – often with inherent uncertainties – while adhering to specific operational needs and constraints. This work analyzes 30 *go/no-go* decisions made by professional analysts, who regularly make such decisions for their respective organizations, to understand their decision-making strategies. These insights are then used to inform the development of predictive models, including both deterministic rule-based classifiers and AI approaches.

A comparative analysis of decision-making practices across five organizations is conducted. Six analysts from these entities make *go/no-go* decisions for a set of 30 critical conjunction events. It is found that the analysts make significantly different decisions when presented with the same events. This disparity arises from the use of varying decision criteria and thresholds. The factors influencing each analyst's decisions are explored, showing that while probability of collision (PoC) is a key factor, some analysts prioritize distance-related features (miss distance/radial separation) more. Trend information is vital, with all six analysts relying on feature evolution (across PoC, relative position, covariance) in their decision-making process. Additionally, there is a notable difference in how analysts handle high uncertainties or irregular trends, with some choosing to hold off on issuing a *go* decision despite a high PoC in the presence of uncertainty. The application of the Dempster-Shafer Theory to combine *go/no-go* decisions is explored in this work. This approach provides a systematic method for integrating decisions from multiple sources, such as analysts in this case. By considering both supporting evidence and uncertainty, it enables the generation of a single, fused decision that reflects the collective input of the analysts, while accounting for variations in their confidence levels.

Additionally, deterministic models are devised to automate decision-making by replicating analysts' decision criteria. Rule-based classification models are developed using insights from the analysts' survey responses and use combinations of thresholds across parameters to classify events. However, they are tedious to tailor to each individual analyst and their rigidity limits generalization. In contrast, a Long Short-Term Memory (LSTM) model is introduced as a more flexible alternative, capable of learning decision patterns directly from data. The model autonomously extracts meaningful relationships between the features in the dataset and outperforms the rule-based classifiers, achieving an accuracy and an F1-score of 88% on a test set of 1600 critical events. Learned feature importance analysis on the LSTM model indicates a heavy reliance of the model on distance-based quantities, such as radial separation, with little influence from typically dominant features in decision-making such as the PoC.

REFERENCES

1. European Space Agency, (2025). *Space debris by the numbers*
2. Orbiting Now, (2024). *Active satellite orbit data*, https://orbit.ing-now.com/
3. Lawrence A., Rawls M. L., Jah M., Boley A., Di Vruno F., Garrington S., et al. (2022). *The case for space environmentalism*, Nature Astronomy, **6**(4), 428–435. DOI: 10.1038/s41550-022-01655-6
4. McKnight, D., Dale, E., Bhatia, R., Patel, M., Kunstadter, C. (2022). *A Map of the Statistical Collision Risk in LEO*, Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS), Maui, HI, USA, September 27-30
5. Virgili, B. B., Flohrer, T., Krag, H., Merz, K., Lemmens, S. (2019). *CREAM - ESA's Proposal for Collision Risk Estimation and Automated Mitigation*, First International Orbital Debris Conference, Sugar Land, TX, USA, December 9-12
6. SpaceX, (2022). *SpaceX's Approach to Space Sustainability and Safety*, https://www.spacex.com/updates/sustainability
7. Uriot T., Izzo D., Simões L., Abay R., Einecke N., Rebhan S., Martinez-Heras J., Letizia F., Siminski J., Merz K., (2022). *Spacecraft collision Avoidance challenge: Design and results of a machine learning competition*, Astrodynamics, **6**, 121–140
8. Abay R., Caldas F., Filipe M., Guimarães M. (2021). *Benchmarking Machine Learning Models for Collision Risk Prediction in Low Earth Orbit*, 8th European Conference on Space Debris, Darmstadt, Germany, April 20–23
9. Schaus V., Andriof T., Borrett C., Burmeister I., Cabral F., Carvalho J., et al. (2022). *First results of*

ESA's collision risk estimation and automated mitigation (CREAM) programme, 73rd International Astronautical Congress (IAC), Paris, France, September 18–22

10. Pinto F., Acciarini G., Metz S., Boufelja S., Kaczmarek S., Merz K., et al. (2020). *Towards Automated Satellite Conjunction Management with Bayesian Deep Learning*, AI for Earth Sciences Workshop, December 12–13

11. Stroe I.F., Stanculescu A.D., Ilioaica P.B., Blaj C.F., Nita M.A., Butu A.F., et al. (2021). *AUTOCA, Autonomous Collision Avoidance System*, 8th European Conference on Space Debris, Darmstadt, Germany, April 20–23

12. Ravi P., Zollo A., Fiedler H., (2023). *AI for Satellite Collision Avoidance – Go/No Go Decision-Making*, Second International Orbital Debris Conference, Sugar Land, TX, USA, December 4-7

13. Guimarães M., Soares C., Manfletti C. (2023). *Predicting the Position Uncertainty at the Time of Closest Approach with Diffusion Models*, 74th International Astronautical Congress (IAC), Baku, Azerbaijan, October 2–6

14. Stevenson E., Rodriguez-Fernandez V., Urrutxua H., Morand V., Camacho D. (2021) *Artificial Intelligence for All vs. All Conjunction Screening*, 8th European Conference on Space Debris, Darmstadt, Germany, April 20–23.

15. Stevenson E., Rodriguez-Fernandez V., Urrutxua H. (2022). *Towards graph-based machine learning for conjunction assessment*, Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS), Maui, HI, USA, September 27-30

16. Henry, S., Armellin, R., Gateau, T. (2023). *Safe-event pruning in spacecraft conjunction management*, Astrodynamics, **7**, 401—413

17. Hernández C.P., Lubián-Arenillas D., Periañez C.P., Velez J.T., Solomon A. (2022). *Should I stay or should I go? Machine Learning applied to Conjunction Analysis*, 73rd International Astronautical Congress (IAC), Paris, France, September 18–22

18. Sánchez L., Vasile M., Minisci E. (2021). *On the use of Machine Learning and Evidence Theory to improve collision risk management*, Acta Astronautica, **181**, 694–706. DOI: 10.1016/j.actaastro.2020.08.004.

19. Salvatore A., Oltrogge D., Arona, L., (2022). *Operators' Requirements for SSA Services*, Journal of the Astronautical Sciences, **69**, 1441–1476.

20. Kerr E., Ortiz N. S., (2021). *State of the art and future needs in conjunction analysis methods, processes and software*, 8th European Conference on Space Debris, Darmstadt, Germany, April 20–23

21. Schiemenz F., Utzmann J., Kayal, H., (2019). *Survey of the operational state of the art in conjunction analysis*, CEAS Space J, **11**, 255–268.

22. Aida S., Kirschner M., (2012). *Operational Results of Conjunction Assessment and Mitigation at German Space Operations Center*, Transactions of the Japanese Society for Aeronautical and Space Sciences, **28**, 23–28

23. Sentz K., Ferson S., (2002). *Combination of Evidence in Dempster-Shafer Theory*, Sandia National Laboratories

24. Hochreiter S., Schmidhuber J., (1997). *Long Short-Term Memory*, Neural Computation, **9**(8), 1735–1780

25. Sokolova M., Lapalme G., (2009). *A systematic analysis of performance measures for classification tasks*, Information Processing and Management, **45**, 427–437