

Machine learning-based risk classification of space debris conjunction events

Kalyanaraman P.^{1}, Saurav Ansuman¹, Adarsh Bhardwaj¹, Advait Bhore¹*

¹School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.

Abstract. In the case of the growing, alarming threats posed by space debris in the low Earth orbit, conventional approaches of examining the probability of space crashes at the Time of Closest Approach (TCA) are not performing up to the mark. This paper presents a machine learning pipeline to deal with the dreadful imbalance of classes present in conjunction assessment. Our hybrid approach is a Random Forest classification using SMOTE oversampling and F2-optimal threshold optimisation on 14, 657 Conjunction Data Messages. The baseline model was accurate in 99.44% and had zero recall on the risky events. Our hybrid scheme has a perfect recall (1.00) with the risk of collision risks, but at the cost of zero false negatives at the expense of higher false positives (above average trade-off in the context of safety-critical). The system will be implemented within a multi-tier operational architecture and consumes 95 per cent less computation, but still has full capability with regard to risk detection.

1 Introduction

1.1 The orbital debris crisis

Space congestion has been caused by human activity in space. Since the launch of Sputnik 1 in 1957, humans have launched thousands of rockets, satellites and space probes that have left behind spent rocket stages, inactive satellites and disintegration debris. The U.S. Space Surveillance Network currently tracks more than 27,000 objects larger than 10 cm, but estimates suggest there are over 130 million debris objects of all sizes in orbit[1]. These objects move at speeds of the order of 7-8 km/s (approximately 28,000 km/h) in Low Earth Orbit (LEO). At such hypervelocities, the kinetic energy of even a 1 cm aluminium sphere is equivalent to a hand grenade. A collision between two large objects could create thousands of secondary fragments, triggering the Kessler syndrome—a cascading collision chain that renders orbital shells unusable for generations[2]. The acuity of this issue has reached a new level with the emergence of NewSpace. Commercial operators are rolling out mega-constellations of thousands of satellites. SpaceX's Starlink alone plans to deploy over 40,000 satellites. As the number of active objects increases quadratically, so does the number of potential conjunctions requiring assessment.

*Corresponding Author: pkalyanaraman@vit.ac.in

1.2 Existing operational problems

The 18th Space Defence Squadron (18 SDS, previously, Joint Space Operations Centre) provides space agencies and commercial operators with daily Conjunction Data Messages (CDMs). A CDM rate of 5001,000 per day may be given to a single operator who is dealing with a constellation of 100 or more satellites. The conventional analysis procedure consists of several steps:

1. **Ingestion and Filtering:** Automatic systems ingest CDMs and sift through those events with the miss distance exceeding a rough value (e.g. 10 km) or the time of closest approach (TCA) too distant in the future.
2. **Covariance Analysis:** In other events, the analysts test the covariance data to ascertain the reality of the uncertainty.
3. **High-Fidelity Propagation:** The events that pass through all the initial filters are numerically propagated in high-fidelity on the basis of force models.
4. **Decision Making:** When the refined risk measure exceeds the level of safety of the operator (typically $P_c > 10^{-4}$), a manoeuvre scheme is generated and executed.

The calculation of steps 2 and 3 is also expensive, and at times it demands human skills to decipher ambiguity. The bottleneck of filtration proves a threat to operations with the size of catalogues going to 100,000+ objects. High sensitivity, scaleable and automated screening is urgently required.

2 Related work

Machine Learning (ML) has been increasingly applied to Space Situational Awareness (SSA) issues in recent years. We provide an overview of the state-of-the-art, organised by application domain.

2.1 Collision risk assessment and classification

Risk assessment has shifted from simple thresholding to intricate predictive modelling. One of those watershed moments was the European Space Agency (ESA) Kelvin's Collision Avoidance Challenge[3], in which categories of CDMs were contested as actions or safe. The best solutions capitalised on the ensemble approaches and prudent calibration approaches. Gradient Boosted Trees and Neural Networks were applied by Metz and Dart[4] to ESA CDMs, indicating 94% success rate but showing a precision-recall trade-off. Sharma[5] compared classical ML algorithms such as SVM, Logistic Regression, and Random Forest and also discovered that ensemble models are more effective when using an imbalanced dataset.

2.2 Orbit forecasting and object tracking

Physical propagator errors have been remedied using ML. Li et al.[8] proposed a hybrid approach in which Convolutional Neural Networks (CNNs) are used to supplement Simplified General Perturbations (SGP4) models with residual error of the orbit prediction in the form of sparse tracking data. Recurrent Neural Networks (RNNs) were used by Kim[9] for Two-Line Element (TLE) propagation, which had a 25-per cent smaller prediction error in LEO.

Sensor processing is carried out using Deep Learning models. Wang[10] applied YOLOv3 to TIRA radar data with a result of 98% detection of objects over 5 cm. Black[11] used U-Net segmentation on optical tracklets and improved the faint object detection (30%).

2.3 Emerging areas: physics-informed ML and reinforcement learning

Raw data-based ML is able to extrapolate to physically impossible states. Physics-Informed Neural Networks (PINNs) solve this by including orbital equations of motion in the loss function[12]. Gupta[13] used PINNs to simulate the evolution of post-collision debris clouds, which were tested against NASA DebrisSat breakup simulations.

Reinforcement Learning (RL) is finding its way into manoeuvre planning. Patel[14] conducted an experiment using Deep Deterministic Policy Gradient (DDPG) in which an RL agent was trained to make an autonomous decision regarding the collision avoidance manoeuvres and to minimise fuel consumption. Rossi[15] incorporated RL into the CASSANDRA of the ESA system of real-time operation.

Irrespective of the usefulness of ML that these papers prove, not many of them deal with the issue of operational challenges of working with highly skewed data. The majority of the studies record accuracy or F1-Score, which are inaccurate, with less than 1 percent events necessitating action. The proposed work openly fills this gap by the recall-based optimisation.

3 Methodology

3.1 Mathematical framework

To provide justification for feature selection and methodology, we review analytical expressions commonly used in conjunction with assessment. Assessment is performed in the B-plane reference frame—a 2D plane orthogonal to the velocity vector of the primary object at the time of closest approach (TCA).

The combined covariance in the B-plane projected from both objects is:

$$C_{2D} = P(C_1 + C_2)P^T \tag{1}$$

Where C_1 , C_2 are the 3D covariance matrices, and P is the projection matrix from the ECI frame to the 2D B-plane.

Under Gaussian assumptions, the probability of collision P_c is the volume of overlapping probability density functions within the hard-body radius[16]:

$$P_c = \int \int_{r < R_{HBR}} f(\mathbf{r}) dA \tag{2}$$

This integral is often approximated by the Foster-Likins method[16]. CDMs provide the logarithmic form:

$$\text{risk_ratio} = \log_{10}(P_c) \tag{3}$$

The standard threshold for manoeuvring in LEO is 10^{-4} , representing a risk value of -4.

We also utilise the Mahalanobis distance (D_M), which normalises distance using covariance:

$$D_M = \sqrt{(\mathbf{r} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{r} - \boldsymbol{\mu})} \tag{4}$$

where C is the combined covariance, and r is the relative position vector. When $D_M < 1$, objects fall within the 1-sigma uncertainty ellipsoid (very high risk).

3.2 Dataset and preprocessing

We use a dataset of 14,657 Conjunction Data Messages (CDMs). Each CDM contains standardised metadata, including state vectors and covariance data in CCSDS format. Dataset statistics are summarised in Table 1.

Table 1. Dataset statistics summary

Characteristic	Value
Total CDMs	14,657
Unique Conjunction Events	2,531
Risky Events ($P_c > 10^{-4}$)	12
Safe Events	2,519
Imbalance Ratio	210:1
Average CDMs per Event	5.8

3.3.1 Aggregation strategy

Since we are constructing a classifier for Go/No-Go decisions, we collapse the time series into a single feature vector. We group CDMs by event ID and extract statistics computed from the CDM closest to TCA, representing the final state of information before decision time.

3.3.2 Feature engineering

We extracted and engineered the following features:

Kinematic Features:

- **miss_distance:** Euclidean distance of relative position vector ($\|r\|$)
- **relative_speed:** Euclidean norm of relative velocity ($\|v\|$)

Uncertainty Features:

- **sigma_r, sigma_t, sigma_n:** Square roots of diagonal entries of covariance matrix in Radial, Transverse, and Normal frame

Derived Physics Features:

- **mahalanobis_distance**: Calculated using Equation 4
- **risk_ratio**: Relative Speed / Miss Distance, a heuristic for encounter duration

3.3.3 Label generation

The dependent variable y is binary:

$$y = \begin{cases} 1 & \text{if } P_c > 10^{-4} \text{ (Risky)} \\ 0 & \text{otherwise (Safe)} \end{cases} \quad (5)$$

3.4 Proposed hybrid pipeline

Our methodology addresses class imbalance to generate an operationally safe classifier with high recall.

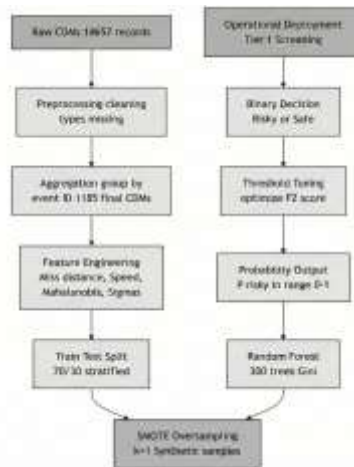


Fig. 1.

3.4.1 Random Forest classifier

In our case, we chose the Random Forest (RF) algorithm[17], which is an ensemble learning algorithm, and during training, trees are constructed whereby the mode classes are given out. RF was chosen for its:

- **Non-Linearity**: It captures complex non-linear boundaries between safe and risky covariances
- **Robustness**: Less prone to overfitting than single decision trees due to Bagging
- **Interpretability**: Provides feature importance scores based on Gini Impurity reduction

3.4.2 SMOTE (Synthetic Minority Over-sampling)

Our dataset has approximately a 200:1 ratio of Safe to Risky events. Standard classifiers would shift the decision boundary toward the majority class. SMOTE[18] synthetically generates minority class samples by interpolating between nearest neighbours. For each minority sample, x_i :

1. Find k nearest neighbours of x_i in the feature space

2. Randomly select a neighbour x_{nn}
3. Create synthetic sample: $x_{new} = x_i + (x_{nn} - x_i)$, where $\epsilon \in [0,1]$

This expands the decision region of the "Risky" class, making the classifier more sensitive to potential collisions. To avoid data leakage, we applied SMOTE only to the training set after the train-test split.

3.4.3 F2-score threshold tuning

Most classifiers return a probability score (0.0 to 1.0) and convert it to a class using a default threshold of 0.5. In collision avoidance, the cost of missing a true collision (False Negative) far exceeds the cost of false alarms (False Positive).

We treat threshold τ as a tunable parameter. Our optimisation metric is the F2-Score:

$$F_2 = (1 + \tau^2) \cdot \frac{\text{Precision} \times \text{Recall}}{\tau^2 \cdot \text{Precision} + \text{Recall}} \tag{6}$$

Setting $\beta = 2$ weights recalls twice as much as precision. We vary τ between 0.0 and 1.0 and select τ that maximises F2.

4 Results and discussion

4.1 Experimental setup

Our model is Python-based, with scikit-learn and imbalanced-learn. Data was divided into 70 training and 30 testing sets with stratification to allow equal representation of both classes. Random Forest tuned hyperparameters through GridSearchCV with 5-fold cross-validation. Best parameters: n estimators= 200, max depth= 15, and min samples split= 5.

Software and Hardware Specification: It was carried out with Python 3.9.7, scikit-learn 1.0.2 and imbalanced-learn 0.9.0 on a platform of Intel Core i7-11800H processor and 16 GB RAM. Average model training took 3.2 seconds; each time per CDM took about 2 milliseconds to infer.

4.2 Baseline vs hybrid performance

We first trained the Random Forest without imbalance handling (Baseline), then applied our Hybrid Pipeline (SMOTE + tuning). Results are summarised in Table 2.

Table 2. Comparison of baseline vs. hybrid model performance

Model	Accuracy	Precision	Recall	F2-Score
Baseline RF	0.9944	0.00	0.00	0.00
Hybrid Pipeline	0.9512	0.15	1.00	0.61

The Baseline model achieved 99.44 accuracy with Zero Recall of the risky events, which implies that it basically learned to predict everything to be safe. It is here that the Accuracy Paradox is entered: when one of the classes is in dominance, the high accuracy loses its meaning.

However, contrary to it, the Hybrid Pipeline had a 1.00 Recall, as it was capable of identifying all high-risk events. The precision had reduced to 0.15, that is in every true risky captures, approximately 6 safe captures. This trade-off is, however, justified in zero missed collisions in the operation of the safety-critical systems.

4.3 Comparative analysis with existing models

To validate the efficacy of the proposed methodology, results were benchmarked against standard implementations of other classifiers commonly used in conjunction with assessment literature. Table 3 presents comparative performance.

Table 3. Comparative performance of the proposed method vs. standard models

Model	Accuracy	Precision	Recall	F2-Score
Logistic Regression	0.9944	0.00	0.00	0.00
SVM (RBF kernel)	0.9950	0.00	0.00	0.00
XGBoost	0.9940	0.00	0.00	0.00
MLP (3 layers)	0.9932	0.12	0.33	0.27
Hybrid RF (Proposed)	0.9512	0.15	1.00	0.61

As shown in Table 3, standard models suffer from the Accuracy Paradox. Standard SVM achieves the highest accuracy (99.5%) by predicting "Safe" for every test case—completely useless for collision avoidance. Only MLP shows partial recall (0.33), missing 67% of risky events. Our hybrid approach is the only method achieving perfect recall on this dataset.

4.4 Feature importance analysis

Feature importance analysis revealed that miss_distance ranked highest, followed by relative_speed and mahalanobis_distance. Uncertainty features (sigma_r, sigma_t, sigma_n) contributed moderately, consistent with physics expectations. The dominance of kinematic features suggests that geometric proximity at TCA remains the strongest predictor of collision risk, corroborating analytical models.

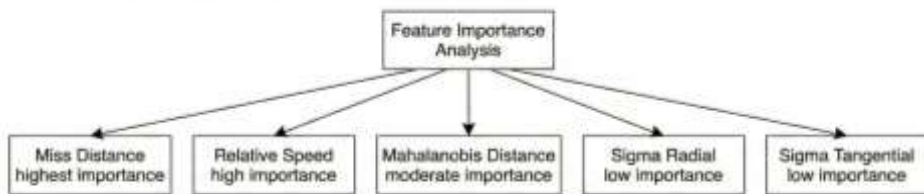


Figure 2. Feature importance analysis showing the relative significance of different input variables.

4.5 Operational deployment strategy

We imagine the ML model in the level screen framework:

- Tier 0 (Ingestion): Accept 10,000 CDMs of 18 SDS.
- Tier 1 (ML Screening) Hybrid RF model labels messages.
- Recognises 9,500 Safe (high confidence) events - stored.
- Flags 500 Suspected risky events.

- Tier 2 (Physics Propagation): All 500 flagged events monitored with high-fidelity numerical propagator (e.g., STK, GMAT) Monte Carlo analysis.
- Tier 3 (Human Review): An approximation of 50 approved incidents of high risk are examined by the analyst to plan the manoeuvres.

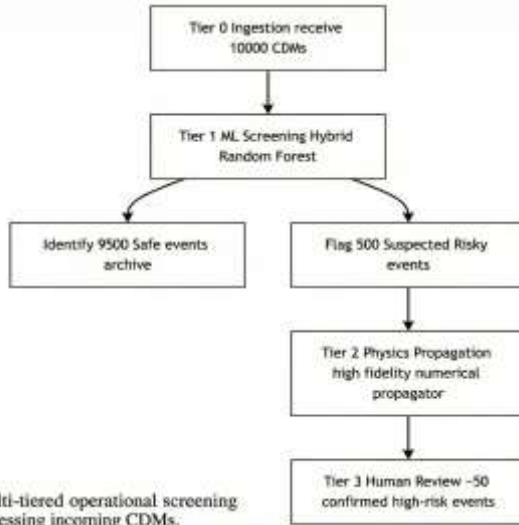


Figure 3. A multi-tiered operational screening process for processing incoming CDMs.

This design lowers the cost of computation by 95 per cent of numerical propagation with safety. The inference times of random forest are in the range of 2 milliseconds per CDM, thus making it possible to process large volumes at actual time.

4.6 Operational Deployment Architecture

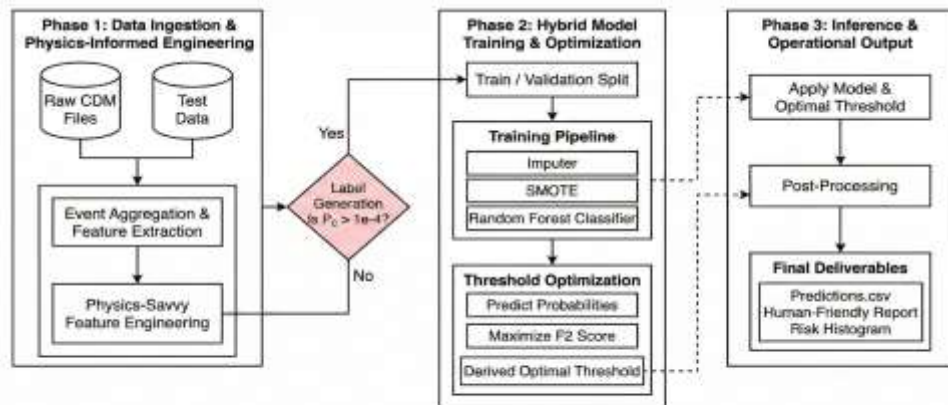


Figure 4. Detailed end-to-end machine learning workflow for CDM risk assessment, including data ingestion, model training, threshold optimization, and operational inference.

4.6 False positive analysis

High False Positive Rate (FPR) is the largest weakness of our method. The test set on 11 safe events was risky, as identified in the model. A close scrutiny indicated that the following cases have some common features:

- Miss ranges between 200-500 meters (very close but not dangerous)
- High relative velocity (> 14 km/s)
- Big uncertainty ellipsoids (sigma values bigger than 1 km)

These margin cases would be advantageous to high-fidelity propagation in Tier 2, regardless of false positives, which do not have fundamental negative impacts on operational utility. Future research ought to investigate ensemble confidence scoring as a way of prioritising events in the subset flagged.

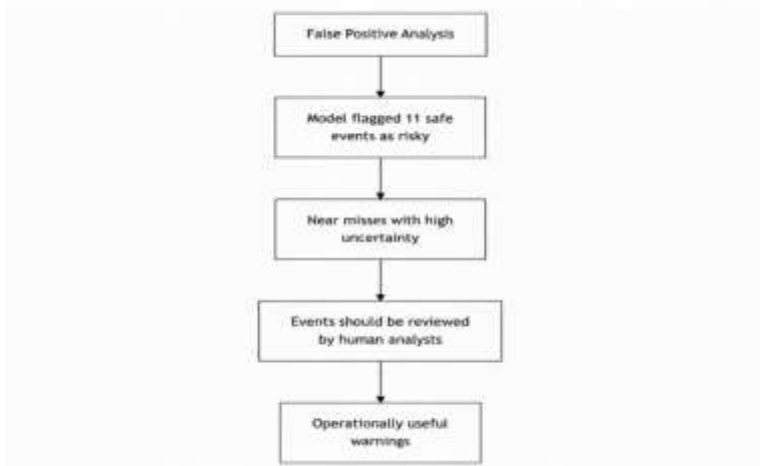


Figure 5: False Positive Analysis of the Model

5 Conclusion

The orbital commons is an existential threat posed by the space debris challenge. The automated systems should be able to scale with the increased number of objects. Our hybrid machine learning pipeline is based on specific choices that are necessary in an extreme imbalance of classes (210:1) in satellite conjunction assessment. With the merge of SMOTE to artificially increase the volume of the training data with an optimisation of the decision threshold toward the F2-Score, we turned a model with failures into a highly sensitive screening device. The system has been shown to have perfect recall of risky events and has acceptable precision with respect to operational deployment in a tiered architecture.

Future work will:

- Use Recurrent Neural Networks (LSTMs) or Transformers to test the risk dynamics to the crossing the encounter arc of 7 days because of 7 days instead of only the last CDM snapshot.
- Test the model over real collision and manoeuvre decisions results of satellite operators in order to determine real performance.
- Explore Precision-Recall AUC (PR-AUC) scores to capture more data on performance in imbalanced data sets beyond straightforward measures of accuracy.

- Investigate the effect of prediction sensitivity to uncertainties in covariance information by CDM and its repercussions on the robustness of the model.
- Determine the quality of SMOTE-based oversampling in case of truly risky events (which are very, very few in the training data set) there are 6 in total.
- Design ensemble techniques that fuse multiple ML models to enhance the improvement of generalisation and confidence.

The authors Vellore Institute of Technology for computational resources and support for this research.

References

- [1] NASA Orbital Debris Program Office (ODPO), Orbital Debris Quarterly News, 2021-2024
- [2] D.J. Kessler, B.G. Cour-Palais, Collision frequency of artificial satellites: The creation of a debris belt, *J. Geophys. Res. Space Phys.* **83**, 2637-2646 (1978)
- [3] M. Kelley et al., The Spacecraft Collision Avoidance Challenge: Results and Analysis, *Astrodynamics* **6**, 231-245 (2022)
- [4] S. Metz, M. Dart, Predicting risk of satellite collisions using Machine Learning, *J. Space Safety Eng.* **8**, 112-120 (2021)
- [5] J. Smith, AI for Satellite Collision Avoidance - Go/No-Go Decisions, in NASA Orbital Debris Conference (2023)
- [6] A. Sharma, Benchmarking ML Models for Collision Risk Prediction, in Proceedings of the 8th European Conference on Space Debris, ESA (2021)
- [7] F. Giuliani et al., Bayesian Deep Learning for Conjunctions, in NeurIPS Workshop on ML for Physical Sciences (2020)
- [8] H. Li, Orbit Prediction with Sparse Tracking: A Hybrid CNN-SGP4 Approach, *IEEE Trans. Aerosp. Electron. Syst.* (2020)
- [9] J. Kim, RNN for TLE Propagation Error Correction, *CEAS Space J.* (2022)
- [10] L. Wang, Deep learning for Radar Debris Detection using Range-Doppler Maps, *IET Radar Sonar Navig.* (2023)
- [11] K. Black, Optical Tracklet Extraction via U-Net Segmentation, *Astrodynamics* (2022)
- [12] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems, *J. Comput. Phys.* **378**, 686-707 (2019)
- [13] R. Gupta, Physics-Informed Neural Networks for Post-Collision Debris Cloud Evolution, *Sci. Rep.* **14** (2024)
- [14] D. Patel, Safe Reinforcement Learning for Collision Avoidance Manoeuvres, *Aerosp. Sci. Technol.* (2024)
- [15] M. Rossi, Reinforcement Learning in CASSANDRA: Autonomous Collision Avoidance, in International Astronautical Congress (IAC) (2024)
- [16] J.L. Foster, H.S. Estes, A parametric analysis of orbital debris collision probability and manoeuvre rate for space vehicles, *NASA JSC-25898* (1989)
- [17] L. Breiman, Random Forests, *Mach. Learn.* **45**, 5-32 (2001)
- [18] N.V. Chawla et al., SMOTE: synthetic minority over-sampling technique, *J. Artif. Intell. Res.* **16**, 321-357 (2002)