Array 30 (2026) 100799

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

A physics-informed machine learning for detecting suspicious satellite maneuvers (orbital manipulation)

K.K.H. Karunathilake [a,*], Kavinga Yapa Abeywardena [b], Sara Vecchini [c]

[a] Faculty of Graduate Studies and Research, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
[b] Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
[c] Security Operations and Data Science, 2T Security Ltd., Leicester, United Kingdom

ARTICLE INFO

Keywords:
Satellite security
Orbital manipulation
Anomaly detection
Cybersecurity
Machine learning

ABSTRACT

—Satellite systems have become prime targets for cyberthreats given their critical role in global infrastructure and general lack of security. Among these, orbital manipulation, a form of satellite hijacking, is a particularly severe threat that can disrupt essential operations and impact national security. To address these concerns, this research proposes an Artificial Intelligence (AI)-based anomaly detection system that utilizes Machine Learning (ML) models to analyze telemetry data for possible orbital manipulations with a multi-gate physics architecture grounded in orbital mechanics, to verify that detected anomalies are kinematically inconsistent and are therefore genuine integrity failures. This research demonstrates that temporal-based models like LSTM are essential for this domain, achieving high recall rates which are then validated by the physics component. While the framework includes multiple physical constraints, this study specifically validates the energy-based Vis-Viva gate, with the Tsiolkovsky and Angular Momentum gates established as architectural designs for future verification. This study concludes that successful AI deployment in orbital cybersecurity requires a comprehensive approach that integrates domain-specific context and physics-informed validation beyond traditional performance metrics.

1. Introduction

With the rapid advancement in technology, satellites are becoming a core component of critical infrastructure around the world. This is because satellites play a vital role in essential services that fortify modern-day economies, national security, and public safety. Key contributions of satellites to critical infrastructure include (but are not limited to) global communications and connectivity, national security and defense, navigation and positioning, and earth observation and climate monitoring [1]. Satellite communication systems consist of three main segments, namely, 1) *the space segment* (that contains one or more active or spare satellites organized into a constellation), 2) *the control segment* (that consists of all ground facilities used for controlling and monitoring the satellites as well as for managing the traffic and associated resources), and 3) *the ground segment* (that consists of all the traffic Earth stations) as shown in Fig. 1 [1]. However, with critical infrastructure increasingly relying on satellite systems, today, they are gradually becoming targets to cyberattacks like jamming, spoofing, and hijacking. For example, in 2022, during Russia's invasion of Ukraine, in efforts to disable Ukrainian military communications, Russia targeted the American satellite organization – Viasat, thereby impacting the organization's satellite operations across Europe [2,3]. Furthermore, when Elon Musk offered Ukraine to access SpaceX's growing Low Earth Orbit (LEO) communication satellites – Starlink, it was soon reported that Starlink too was suffering from jamming and hacking attempts from Russia [2]. Moreover, in 2022, a hacktivist group by the name Network Battalion 65 (NB65), asserted that they managed to successfully breach the satellite imaging capabilities of the Russian State Corporation for Space Activities (ROSCOSMOS) as a response to Russia's invasion of Ukraine [4]. Although these claims were denied by ROSCOSMOS claiming that their systems were continuing to operate as expected, these events from recent history are prime examples to highlight how the threat landscape is evolving today to target satellite systems and inflict more damage to critical infrastructure across the world. These case studies also highlight the increasing need to ensure the security of satellites to provide confidentiality, integrity, and seamless availability of their services. As such, among the numerous threats to satellites, this research will focus on satellite hijacking.

Satellite hijacking occurs when threat actors gain unauthorized access to satellite systems. This can lead to the alteration or disruption of

* Corresponding author.
E-mail addresses: ms24014922@my.sliit.lk (K.K.H. Karunathilake), kavinga.y@sliit.lk (K.Y. Abeywardena), sara.vecchini@2t-security.com (S. Vecchini).

https://doi.org/10.1016/j.array.2026.100799
Received 23 October 2025; Received in revised form 8 March 2026; Accepted 31 March 2026
Available online 1 April 2026

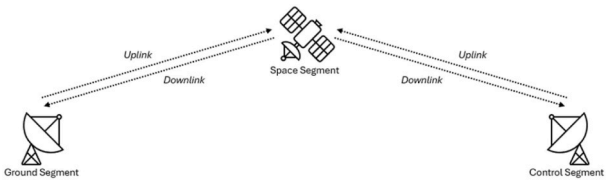

Fig. 1. Components of a satellite communication system.

data transmissions or even losing control of the satellite itself. As such, by disrupting confidentiality, integrity, and availability of critical services, satellite hijacking poses a threat not only to aerospace corporations, but also to national security.

Furthermore, it can be seen that the absence of fundamental security controls in satellite systems exposes them to various threats including satellite hijacking. Satellite hijacking has many security implications including military and intelligence disruption, cyberwarfare and espionage, internet and telecommunication blackouts, civil aviation and maritime risks, and even Anti-Satellite (ASAT) attacks. There are several ways through which satellite hijacking can occur including (but not limited to) command spoofing, radio frequency jamming, and replay attacks. Although radio frequency jamming is the most common form of satellite hijacking, these attacks mostly target media broadcasts of satellite television signals where the uplink signal is replaced with a stronger and malicious radio transmission [5]. This is mainly because, although Global Navigation Satellite System (GNSS) signals are of critical nature, they could be spoofed with ease using cheaper and common tools due to the signals themselves being weak by the time they reach the Earth [5]. Satellite hijacking also tends to have drastic implications, especially for LEO satellites that function in large numbers [6]. For example, in a LEO constellation, if one satellite has a positioning error, it could disrupt the entire constellation and could also pose threats of collision with other satellites [6].

In a recent publication, Willbold et al. surveyed 19 engineers representing 17 different satellite models [7]. Out of these 17 models, three (i.e., 17.65%) were found to not have implemented any security measure to prevent intrusions whereas engineers of five out of the 17 satellite models were unsure of any security implementations or failed to comment (i.e., 29.41%) [7]. However, the remaining nine had some level of security implementations done (i.e., 52.94%), although the level of security provided by these implementations remain questionable [7]. This is because, even among the nine satellite models with some level of security implementations in place, only five had implemented any form of access controls [7]. The outcome of this survey is especially alarming as satellites are continuing to become an integral component of the technological world today.

The need for extensive research in this domain can be further justified by referring to Jones, who in her publication, explored the perspective of space professionals on satellite security by comparing the results of two similar surveys conducted in 2012 and 2022, with a decade of evolution in between the two surveys [8]. According to the author's findings, within a decade, cyberattacks on satellite systems have emerged to be within the top three risks to satellites operating in outer space [8]. When comparing the likelihood of gaining unauthorized access to satellite data, the Likert Scale indicators for positive groupings notably increased from 47.00% to 66.00% from 2012 to 2022 [8]. Furthermore, when comparing the findings of the possibility of satellite hijacking between 2012 and 2022, the positive groupings in the Likert Scale of the survey significantly increased from 21.00% to 39.00% indicating the growing likelihood of satellite hijacking attempts in the world today [8]. As such, with satellite hijacking emerging as a common threat to satellite systems, identifying an effective solution to address this threat has become crucial. As satellite hijacking by itself is a broad domain, this research will focus on designing a solution capable of identifying anomalies in satellite telemetry data to detect a type of satellite hijacking attacks – satellite orbital manipulation attacks.

Satellite orbital manipulations attacks are a form of satellite hijacking attacks where an adversary could alter the orbit orientation of a satellite by gaining unauthorized control over a satellite's systems like its propulsion and attitude control. Furthermore, satellites are perfect examples of Cyber Physical Systems (CPS) as they integrate physical processes with computation and networking. As such, a successful cyberattack on a CPS would have tangible physical consequences and therefore, manipulating the orbit of a satellite which is a physical process is a direct result of a successful cyberattack on the digital systems that control it.

A successful satellite orbital manipulation attack would have several security implications including (but not limited to) destruction and debris, service disruption, espionage and misdirection, weaponization, threat to national security, violation of privacy, commercial and economic risks, fake news and propaganda, and cyberwarfare and space conflicts.

For example, in October 2021, China launched a satellite called “Shijian-21 (SJ-21)” into orbit [9]. Although its operational details were classified, a commercial space monitoring company called ExoAnalytic Solutions claimed that on January 22nd, 2022, SJ-21 went absent from its orbital slot for several hours and instead approached the defunct CompassG2 satellite to dock and drastically alter its geostationary orbit, pulling it into a space graveyard [9]. Although China previously claimed that SJ-21 would mainly be used to test and verify space debris mitigation strategies, their lack of transparency compared to the efforts of other countries raised serious doubts regarding the military prospects of such dual technologies and the threat they could thereby impose [9]. This is a prime example to show the growing concerns around cyberwarfare and space warfare in the world today and it further justifies the requirement for extensive research in this domain.

This research study is expected to make several key contributions to both academic knowledge and practical applications as follows.

1. Establish a foundational understanding of satellite orbital manipulation, analyzing real-world attack vectors and existing cybersecurity gaps to inform the development of a ground-based threat detection architecture.
2. Design and implement a novel AI-based anomaly detection model specifically engineered to identify satellite orbital manipulation attempts by fusing data-driven learning with fundamental engineering principles.
3. Integrate the laws of orbital mechanics as a non-negotiable, physics-informed validation constraint within the anomaly detection pipeline of the ground station, enhancing model robustness and virtually eliminating false positives related to telemetry drift.
4. Rigorously validate the developed model to demonstrate superior threat detection capabilities and successfully achieve a target recall of at least 85.00% in identifying orbital manipulation attempts.
5. Provide a full implementation and evaluation of the end-to-end solution, offering a detailed assessment of its performance and demonstrating the essential synergy between its AI and physics-based components through dedicated validation to help establish AI-driven security standards for satellite systems.
6. Bridge a significant gap between AI research and practical cybersecurity efforts by translating advanced AI anomaly detection from theoretical modeling to a validated, implementable system.

This research paper is divided into nine sections: Section I introduces the reader to the research, Section II provides a literature review, and Section III discusses the research problem. Furthermore, Section IV details the design and implementation of the research followed by the results and discussions under Section V. Section VI discusses the limitations of this research and Section VII entails prospects for future research before concluding the research under Section VIII.

2. Literature review

For years now, ransomware has continued to be one of the formidable threats to systems worldwide. In their publication, Falco et al. explored the possibility of deploying a ransomware attack on an aerial vehicle that engages with the core Flight System (cFS) developed by National Aeronautics and Space Administration (NASA), and has been used for many space and satellite missions [10]. In the context of satellite systems, this publication highlights the possibility of hijacking a satellite while paving a way to restore access to the owner if a ransom is paid [10]. The authors also highlight how with business models like Ransomware-as-a-Service (RaaS) at large today, aerospace corporations that do not prioritize security as an integral component, are only one tailored ransomware kit from RaaS developers away from losing control of entire satellite constellations [10].

In their publication, Benecki et al. introduced a Genetic Algorithm (GA) that improves the anomaly detection capabilities of satellite telemetry using evolutionary thresholding and Long Short-Term Memory (LSTM) nets [11]. According to the authors, there are three types of anomalies that need to be considered for complex systems like satellite systems, namely, 1) *in-point anomalies* (the telemetry values that deviate from the nominal operational range), 2) *collective anomalies* (the overall consecutive telemetry data that are anomalous), and 3) *contextual anomalies* (singular telemetry data that are anomalous within their neighborhood) [11]. These anomaly detection models often deteriorate in performance due to incorrect hyperparameters and can be corrected using the proposed GA [11].

Jin et al., in their publication, introduced a cluster-based method for anomaly detection of satellite telemetry data under unsupervised conditions [12]. Initially, the proposed model clusters the satellite telemetry data with arbitrary shapes using an extended dominant sets clustering algorithm, which then classifies anomalies based on whether an object belongs to a smaller cluster or no cluster at all [12]. Furthermore, this model identifies anomalies in larger clusters using relative similarity and identifies anomaly windows in telemetry sequence based on a local anomaly rate [12]. This model can be considered as an effective model as it improved its Area Under the Curve (AUC) values by 3.00% to 10.00% using the concept of relative similarity, thereby significantly reducing false positives in anomaly detection with further verification using the telemetry data of Tianping-2B satellite [12].

Baireddy et al. introduced an approach to train a general time-series predictor model that could adapt itself to detect anomalies in any telemetry channel, in their publication [13]. This approach consisted of three main steps, namely, 1) *time-series prediction* that utilizes the Recurrent Neural Network (RNN) deep learning model LSTM, 2) *transfer learning* where a single anomaly detection model is trained and similar sub-models inherit information from data in the early layers, and 3) *anomaly extraction* where Kernel Density Estimation (KDE) is used for the prediction error probability distribution function estimation [13]. The various models trained in this publication undergoes an inverse relationship between recall and precision, where the improvement of one metric makes the other drop slightly [13]. The authors also highlight that freezing certain LSTM layers for tuning efforts is a viable strategy where training time is a crucial factor [13].

As opposed to the use of Out-Of-Limit (OOL) method for anomaly detection, Wang et al. proposed a data-driven anomaly detection model that uses Deviation Divide Mean over Neighbors (DDMN) method to address fake anomalies caused by errors in continuous and discrete variables in satellite telemetry data [14]. This model then uses LSTM to model multivariable time series data followed by a Gaussian model that can detect anomalies [14]. This publication further emphasizes three types of data that could lead to false positives in data-driven anomaly detection, namely, 1) *fake anomalies* (that inevitably lead to false positives, but can be finetuned using engineering approaches such as the multi-decision method), 2) *unknown incidents* (that can be caused by certain unprecedented events in the space or the satellite itself), and 3) *sparse samples* (generated when the satellite changes from one mode to another leading to infrequent samples, but can be resolved using error smoothing methods) [14]. When comparing the results of the proposed model with a simple LSTM model at various thresholds, it can be seen that while the precisions were the same for both models, the DDMN model performed exceptionally better than the LSTM model in terms of recall and F1-Score [14].

He et al. introduced a novel Sparse Feature-based Anomaly Detection (SFAD) to identify anomalies in satellite telemetry data, in their publication [15]. Initially, this model obtains a telemetry data dictionary and its corresponding sparse matrix with the use of K-Means Singular Value Decomposition (K-SVD) and defines the sparse features using the sparse matrix [15]. Then, for anomaly detection, lower-dimensional sparse feature vectors are fed into a One-Class Support Vector Machine (OCSVM) model [15]. This publication further explains two approaches that are typically used for spacecraft anomaly detection, namely, 1) *error-based methods* (where a reconstruction model based on training data determines anomalies if the reconstruction errors surpass a predefined threshold), and 2) *similarity-based methods* (that consider samples with low similarity as anomalous from the sample set) [15]. According to the authors, SFAD is a better alternative for error-based and similarity-based models as it is capable of correlating anomalies between various telemetry parameters [15].

Kricheff et al., in their publication, introduced an explainable model for anomaly detection in satellite telemetry by incorporating explainability into Machine Learning (ML) [16]. This approach focuses on two ML models, namely, 1) *LSTM model*, and 2) *hybrid Isolation Forest and Clustering-Based Local Outlier Factor (IF-CBLOF) model*, and then evaluates three explainability models, namely, 1) *Shapley Addictive Explanations (SHAP)*, 2) *Local Interpretable Model-Agnostic Explanations (LIME)*, and 3) *Layer-wise Relevance Propagation (LRP)* [16]. When considering the outcomes of these models, the authors highlight that although SHAP did not meet expectations when met with larger time gaps in the dataset, it performed well when applied individually to IF-CBLOF [16]. The authors further explain that LIME has less accurate and less intuitive outputs when compared to other models and LRP can be successfully leveraged to calculate relevance scores providing more insight into the predictions made by LSTM [16]. As such, the proposed approach utilizes SHAP as the primary explainability model for the ML model IF-CBLOF [16].

Habib et al. in their publication, explored various ML techniques that could be leveraged to identify and predict anomalies in satellite telemetry data, and challenges associated with them [17]. The authors also present another type of anomaly observed in satellite telemetry data in addition to those mentioned by Benecki et al. in their publication called "correlation anomalies" that refer to unprecedented relationships between a multitude of data points [11,17]. In their study, the authors proposed the use of Isolation Forest (IF) for anomaly detection and Random Forest for anomaly prediction, both of which showcased promising results in the initial version of the application [17]. The anomaly prediction component of this research study can be considered as a novel solution that takes proactive security of satellite systems into account [17].

To address issues like high false negatives, longer anomaly detection time, and poor comprehensibility, Zeng et al. proposed a novel anomaly detection mechanism that uses Parametric Causality to identify hidden relationships between spacecraft telemetry data and Double-Criteria Drift Streaming Peaks Over Threshold (DCDSPOT) that improves anomaly detection [18]. The results of this study show that, when compared to traditional statistical models for anomaly detection, the proposed approach detected anomalies with a higher accuracy, recall and F1-Scores [18]. The authors also state that the use of Probability Weighted Moment (PWM) over Maximum Likelihood Estimation (MLE) significantly improves the time consumption problem caused by DCDSPOT as it is not only more computationally efficient, but is also more suited for extreme value distributions like anomalies in spacecraft

telemetry data [18].

Ali proposed a real-time anomaly detection approach for satellite telemetry data using time-series analysis in his publication [19]. The author analyzes four models, namely, 1) *Auto Regressive Integrated Moving Average (ARIMA)*, 2) *Prophet*, 3) *LSTM*, and 4) *Autoencoders* using a dataset in NASA's LEO constellation that contains data like battery temperature, spacecraft bus current data, bus voltage data, and reaction wheel Revolutions Per Minute (RPM) [19]. The proposed system receives the data and then provides feedback to the Flight Control Center (FCC) [19]. When considering the findings of this experiment, the author mentions that ARIMA generalizes well and is a powerful option for forecasting although it is extremely resource intensive and requires hyperparameter tuning, calling for supervision and high capability hardware components [19]. According to the author, the LSTM and Autoencoder models showed overfitting and were overly sensitive to outliers [19]. However, the Prophet model managed outliers well and was robust to missing data and changes in trends, showcasing its superior predictive capabilities [19]. Furthermore, the Prophet model was also very fast compared to the ARIMA model and was less sensitive to outliers compared to its LSTM and Autoencoder counterparts [19].

In their publication, Bieber et al. proposed a generic diagnostic framework for anomaly detection in satellite and spacecraft systems using the proximity-based method K-Nearest Neighbors (KNN), ensemble method Isolation Forest, domain-based method OCSVM, and subspace-based method Principal Component Analysis (PCA) [20]. This approach was validated against three datasets, namely, 1) *NASA Soil Moisture Active Passive (SMAP) satellite dataset*, 2) *Mars Science Laboratory (MSL) rover dataset*, and 3) *European Space Agency (ESA) Reaction Wheels Dataset (RWL) satellite dataset* [20]. The performance of this model is evaluated using three metrics, namely, 1) *F1-Score*, 2) *F1 Point Adjust (F1PA)*, and 3) *F1-Score Composite (FC)* [20]. The authors mention that the quality of the models depend significantly on the thresholding techniques used in model training [20]. When considering the results for the three datasets, it was observed that for the SMAP dataset, the KNN model performed the best whereas for the MSL dataset, the results were unclear as both the Isolation Forest model and the KNN model could be considered good performers [20]. As for the RWL dataset, the authors specified that it was difficult to determine the best model due to the observation of a considerably bigger Pareto front [20].

Unmanned Aerial Vehicles (UAVs), commonly known as “drones”, are aircrafts without human pilots, often controlled by the Ground Control Center (GCC) or by an onboard program, similar to satellites [21]. In their publication, Hu et al. introduced a novel approach for anomaly detection in UAVs leveraging a One-Class Kernel Extreme Learning Machine (OCKELM) as the base model which was trained only on normal data [21]. Next, a Triangular Global Alignment Kernel (TGAK) function was used to replace the Radial Basis Function (RBF) to improve the accuracy of the OCKELM model, resulting in a Triangular Global Alignment Kernel One-Class Extreme Learning Machine (TGA-K-OCELM) model [21]. Finally, Fast Independent Component Analysis (FastICA) was used to manually extract the data features of the Air Laboratory Failure and Anomaly (ALFA) dataset which is a flight log generated by a fixed wing drone performing circular flight over an airport in Pittsburg, USA, for validation [21]. When considering the results of this experiment, the authors indicate that the proposed TGAK-OCELM algorithm with FastICA was able to effectively identify anomalies pertaining to engine failures, aileron failures, and elevator failures, when considering the F1-Score as the primary metric [21].

To address problems associated with data-driven methods like the insufficient availability of anomalous training data and the ineffectiveness in reliably extracting strong inter-parameter spatial-temporal dependencies, the authors Chen et al. proposed an unsupervised anomaly detection approach that uses Causality Enhanced Graph (CEG) neural networks [22]. This approach composed of two steps for feature characterization, namely, 1) *creation of node features* (by leveraging the intrinsic characteristics of the flight data and the temporal features derived from trend decomposition) and 2) *characterization of edge features* (by employing matrix-tensor fusion to combine the causality and attention weights) [22]. The authors also proposed a training model which uses low-rank regularization, which is specifically adapted to the network structure, alongside a flexible smoothing strategy [22]. This model proposed by the authors was also validated against the ALFA dataset which showcased exceptional performance across the precision, recall, and F1-Score metrics [22]. The ablation study revealed that the absence of either of the causality identification components or the attention mechanism from the spatial feature extraction module resulted in performance degradation, and that the causality-based method far surpassed correlation in terms of performance [22]. The ablation study also revealed that although LSTM with Gated Recurrent Unit (GRU) performed well when compared to the temporal feature extraction module in the proposed approach, its generalizability appeared insufficient [22]. The threshold comparison section of the research showcased that there was a significant non-linear relationship between the threshold settings and the model performance, which meant that when the threshold was set too low, the number of false positives would increase, resulting in a sharp drop in recall which in turn results in a lower F1-Score, and vice versa [22]. However, the proposed model was able to mitigate this issue via its adaptive bidirectional smoothing strategy [22].

In their publication, Chen et al. proposed a Convolutional Autoencoder (CAE) and Support Vector Data Description (SVDD) based mechanism for UAV anomaly detection that also uses 0/1 soft-margin loss [23]. The proposed approach combines a One-Dimensional Convolutional Neural Network (1DCNN) with an Autoencoder for feature extraction and reconstruction to efficiently extract the spatial-temporal features of the UAV flight data and a non-linear SVDD with 0/1 soft-margin loss and Bregman Alternating Direction Method of Multipliers (BADMM) to address the non-convex nature of the data [22]. This approach leveraged the ALFA dataset for training purposes and the experimental results indicated that this proposed approach outperformed various models like SVDD and Autoencoder in terms of various metrics like accuracy, precision, recall, F1-Score, and G-Mean [22]. However, the authors also point out that this approach was trained only on healthy data and future improvement efforts would require the introduction of anomalous data into training [22].

Bell et al., in their publication, proposed a stacked recurrent Autoencoder method with dynamic thresholding for anomaly detection in UAV sensor data [24]. The proposed system incorporates an LSTM deep learning Autoencoder with dynamic thresholding and weighted loss function for anomaly detection in the ALFA dataset [24]. The performance of this proposed approach was evaluated against three model variants, namely, 1) *LSTM Autoencoder with static thresholding*, 2) *LSTM Autoencoder with dynamic thresholding*, and 3) *LSTM Autoencoder with dynamic thresholding and dynamic weighted loss function* [24]. The experimental results of this research indicate that the best results for the recall and detection delay were observed for the LSTM Autoencoder model with dynamic thresholding whereas the best performance for the precision and accuracy metrics were observed for the LSTM Autoencoder model with dynamic thresholding and dynamic weighted loss function [24]. The authors highlight that since both dynamic thresholding and dynamic weighted loss function are both versatile in nature, their application on various other ML architectures would garner similar results [24].

In their publication, Chen et al. proposed a causal structure learning approach for hierarchical anomaly detection of on-orbit satellites consisting of two modules, namely, 1) *causal structure learning* and 2) *hierarchical anomaly detection* [25]. The causal structure learning module leverages a graph Autoencoder method based on Graph Neural Network (GNN) to automatically learn the Spatial-Temporal Causal Graph (STCG) of telemetry parameters whereas the hierarchical anomaly detection module leverages a Dynamic Detection Threshold Setting (DDTS) method for dynamic definition of the detection threshold [25]. Furthermore, this approach proposed by the authors also introduced an

STGC-based method for false negative identification as well as an additional method to determine the severity of the satellite anomalies [25]. Validation of this proposed approach was conducted using three publicly available datasets, namely, 1) *MSL dataset*, 2) *SMAP dataset*, and 3) *Key Performance Indicator (KPI) dataset* [25]. However, the authors point out that these three datasets are comparatively "trivial" when compared to the European Space Agency Anomaly Detection Benchmark (ESA-ADB), and validation against this dataset will be done in the future [25]. Further validation of this approach was conducted against the Satellite Attitude and Control System (SACS) dataset as well [25]. When considering the performance of this approach, it was seen that this approach was efficient in false negative reduction [25].

To address various issues in satellite telemetry datasets designed for anomaly detection like unbalanced sample distributions, inconspicuous anomaly characteristics, and scarce fault samples, Cui et al. proposed a Dynamic Time Warping (DTW) method that leverages improved KNN models [26]. Initially, oversampling of small sample time series data is achieved through DTW, after which the variation characteristics are analyzed [26]. Then, DTW is once again used to achieve oversampling of small time series data and Fast-DTW is subsequently used to calculate the difference between different time series data records [26]. Finally, an optimization mechanism is added for a KNN model that uses Fast-DTW for clustering analysis [26]. This approach is evaluated using the Advanced Diagnostics And Prognostics Testbed-Lite (ADAPT-Lite) Electrical Power System (EPS) dataset from NASA [26]. The results of this approach conveyed that DTW was able to successfully address few faulty samples as well as the uneven sample distribution whereas Fast-DTW was able to improve the computational speed while characterizing more features [26]. The authors also indicate that while this approach showcased high accuracy in anomaly detection with faulty samples and unbalanced sample distribution, this method has poor impact in datasets with extreme uneven distributions as well as it requires longer computational time [26].

Song et al., in their publication, introduced an anomaly detection method based on the improved Newman community divisions method, specifically to address the rarity of anomalous data samples [27]. Initially, K-Dynamic Time Wrapping Distance Adjacent Nodes (K-DTWAN) method is used to build a complex network model that is capable of capturing the overall relevance of the anomalous dataset [27]. Based on this model, the Newman community divisions algorithm is improved using the Quantum-behaved Particle Swarm Optimization (QPSO) [27]. This experiment was validated using the ADAPT-Lite EPS dataset and the findings indicated that this approach was able to efficiently identify community structures within the network, thereby enabling accurate classification of anomalous data [27]. The accuracy of this proposed method achieved an average of 96.43% and an F1-Score of 97.76%, and the ablation studies conducted confirmed the validity and the effectiveness of the proposed methodology [27].

To learn generalized normal patterns to then facilitate anomaly detection in spacecraft time series data, Yang et al., in their publication, proposed a Multi-Task Anomaly Detection (MTAD) approach that leveraged unsupervised learning [28]. Initially, four proxy tasks were implemented, namely, 1) *LSTM-based anomaly detection*, 2) *Autoencoder-based latent representation and data reconstruction*, 3) *Variational Autoencoder (VAE)-based latent representation and data reconstruction*, and 4) *joint latent representation-based prediction* [28]. The proxy tasks 1) and 4) were focused on capturing the temporal correlation of the data whereas the proxy tasks 2) and 3) were focused on capturing the spatial correlation of the data, after which Isolation Forest algorithm was used for anomaly detection [28]. When considering the overall performances of the models MTAD, LSTM Autoencoder, Isolation Forest, and VAE, it can be seen that for both experimented contaminations (0.0050 and 0.0020), the MTAD model performed the best across the metrics precision, recall, and F1-Score, with The LSTM Autoencoder coming close behind at the second best overall performance [28]. The ablation study results revealed that the MTAD method was the most successful due to its joint training leveraging all proxy tasks [28].

In their publication, Cuéllar et al. proposed an explainable anomaly detection mechanism for spacecraft telemetry using two datasets from NASA, namely, SMAP and MSL [29]. The authors specify five types of errors that their proposed model can capture, namely, 1) *Type I* (point anomalies which are individual data points in a time series that fall outside the expected range and can be detected by measuring their distance from neighboring values), 2) *Type II* (collective anomalies which are characterized by a significant magnitude variation over an interval in periodic or continuous time series, without notable changes in waveform or frequency and are detectable by weighing the data points within that defined neighborhood), 3) *Type III* (a collective anomaly that occurs when a segment of a periodic or continuous time series exhibits a significant alteration in its waveform even if the absolute amplitude values remain within the normal range and is detectable by calculating the difference between the observed and expected time series values), 4) *Type IV* (a contextual anomaly that is characterized by a significant change in frequency within an interval, while the time-series values themselves are maintained and is detectable by comparing the spectral components across the affected time period), and 5) *Type V* (a contextual anomaly that is characterized by an abrupt variation in the pattern or magnitude trend of continuous time series values, even though the individual data points themselves are not anomalous and is detected by calculating the prediction error between the observed and expected pattern) [29]. The proposed methodology of this research works in two main stages, namely, 1) *feature extraction process* (generates a feature vector containing knowledge to differentiate anomalous behavior) and 2) *training of the ML model* (using the specified features) [29]. The proposed approach obtained the highest average results of precision (95.30%), recall (100.00%), and F0.5-Score (96.20%) when compared to various other models [29].

3. Research problem

Recent literature indicates that anomaly detection in satellite telemetry is a growing domain of interest. However, existing literature largely focuses on anomaly detection itself rather than its application to solve practical, high-stakes security issues like satellite orbital manipulation or hijacking. Furthermore, no established methodology in this domain leverages fundamental orbital mechanics laws to validate ML-generated alerts. This creates a critical gap because ML-only detection suffers from unacceptably high false positives due to its inability to distinguish physical noise from a genuine attack, making such systems unsuitable for operational control centers. The scarcity of publicly available satellite telemetry datasets further hinders the development and validation of robust intrusion detection models. This research is therefore an attempt to address this crucial gap by proposing a novel hybrid methodology. This pipeline uses the temporal strength of ML for statistical anomaly identification and enforces a non-negotiable physics validation gate to ensure that all flagged events represent a true, kinematically inconsistent violation which is the essential kinematic signature of a manipulation attack.

This research makes use of a dataset made public by the ESA called the "ESA-ADB" dataset. This is a large-scale dataset that has real-life satellite telemetry data with curated annotations of anomalies and rare events from three ESA missions and has a total of nine years' worth of data [30]. The selection of this dataset can be further justified using the publication by the authors Chen et al., who highlighted the importance of using this dataset when compared to the other publicly available datasets like MSL, SMAP, and KPI, due to its large-scale, curated, structured, and ML-based nature [31].

To guide this research study, the following hypotheses have been formulated.

H1. A physics-informed, AI-driven model can effectively detect a high proportion of anomalous orbital maneuvers, serving as a reliable

method for identifying the kinematic footprint of a potential satellite orbital manipulation attack.

H2. An AI-based anomaly detection model is more effective than traditional rule-based methods at identifying subtle or time-linked anomalies in satellite orbital behavior.

H3. The integration of physics-based filters into an AI-driven model will result in a more robust anomaly detection system with a low False Negative Rate (FNR) and a balanced False Positive Rate (FPR) when applied to orbital maneuver data.

H4. The inclusion of physics-based features (e.g., changes in orbital energy or specific angular momentum) significantly improves the anomaly detection performance compared to models using raw telemetry data alone.

H5. The change in specific orbital energy for a physically consistent orbital state must fall within an adaptive tolerance ($\leq$51,000.00 J/kg), effectively separating normal LEO dynamics and sensor noise from cyberattack-induced energy discrepancies.

H6. Any instantaneous change in velocity reported by the telemetry must be less than or equal to the maximum physical thrust capability of the satellite ($\leq$250.00 m/s), thereby allowing the Tsiolkovsky Rocket Equation gate to flag kinematically inconsistent changes in velocity attacks.

H7. The orbital plane of the satellite must be conserved; any change in the specific angular momentum vector resulting in a plane change greater than the defined tolerance (1.00°) signifies a kinematically inconsistent state and a violation of the Law of Conservation of Angular Momentum.

H8. For the purpose of high-confidence anomaly detection, the kinematic footprint resulting from an attack-induced orbital maneuver is assumed to be orders of magnitude greater than the accumulated change resulting from non-command perturbations (e.g., atmospheric drag, solar radiation pressure, third-body gravity, or minor debris impact, etc.), allowing the physics-informed filter to effectively distinguish attack signatures from natural orbital dynamics.

H9. Given the proprietary and classified nature of actual satellite cyberattack data, anomalous orbital maneuvers linked to non-malicious telecommand executions within the filtered dataset are a valid and reproducible proxy for the kinematic and spatial-temporal footprint of a satellite orbital manipulation attack, allowing for the successful demonstration of attack detection capability.

4. Design and implementation

In his publication, Aderinto discussed various advanced cybersecurity frameworks which can be used to protect satellite systems, deep-space communication systems, and space assets [32]. The author mentions that among the many threats targeting space systems, AI-powered cyberattacks, quantum computing capabilities, and increase in militarization of the cyberspace can be considered as emerging threats [32]. As such, the publication specifies that the use of AI and ML are becoming indispensable tools in fortifying the security of deep space missions, especially given how AI-driven security solutions provide a proactive approach to anomaly and threat detection, as well as autonomous decision-making, especially in constrained environments [32]. In terms of anomaly detection, the author specifies that AI-driven Intrusion Detection Systems (IDSs), offer flexible self-learning mechanisms which are capable of real-time attack detection and mitigation [32]. The author also specifies that these AI-driven systems are so much more efficient than traditional cybersecurity measures that rely solely on predefined threat signatures [32]. As such, this research focuses on developing an AI-driven, physics-informed anomaly detection system for satellite telemetry that achieves high-fidelity data integrity verification by employing orbital mechanics principles as non-negotiable validation constraints. Fig. 2 presents the physics-informed anomaly detection pipeline at the core of this research, demonstrating the sequential flow from raw telemetry data through data preprocessing and model prediction, culminating in a physics-validated alerting capability.

In his publication, Gregory describes various ways in which satellite-to-satellite attacks could occur, proposing both defense and resilience techniques as well as policies for risk management [31]. All satellite-to-satellite attacks discussed in this publication refer to cyber-attacks, specifically targeting sensors and actuators that facilitate satellite mission capabilities and would have cyber physical consequences [31]. The cyberattack mechanisms discussed in this publication cover situational awareness sensors, electromagnetic pulse actuators, radio frequency actuators, and ground stations [31]. Since processing constraints require satellite decision systems to reside in the ground station, they become prime targets for cyberattacks [31]. As such, an attacker could use a multitude of AI techniques to load a series of processes required to carry out an attack which would then be sent to the satellite for execution via the relevant actuators [31]. The author highlights that due to the lack of constraints on ground stations, they could afford to employ mechanisms that detect and prevent attacks while they are occurring to prolong the life of the satellite [31]. The requirement for centralizing complex detection logic due to On-Board Computer System (OBCS) constraints as defined by the author, can be met through the deployment of the physics-informed anomaly detection model proposed in this work, at the ground station.

Beyond its standalone capabilities, this research is designed to be a crucial component of a broader defense-in-depth cybersecurity framework. The proposed AI system can operate in parallel with existing on-board and on-ground detection tools, providing an independent layer of verification. By functioning as a behavior-based detector, this model offers an external perspective on a satellite's physical state, complementing traditional network-centric security measures that are already in use. As such, this approach also allows for a cross-validation of security events, enabling a single detection from one system to be verified by another, thereby reducing false positives, and significantly strengthening the overall security posture of both individual satellites and entire constellations. Fig. 3 depicts how this cross validation occurs.

Fig. 4 illustrates the proposed hybrid system architecture for detecting satellite orbital manipulation attacks. The workflow begins in the Data Layer, where the raw telemetry dataset undergoes preprocessing and feature engineering to produce the engineered dataset. In

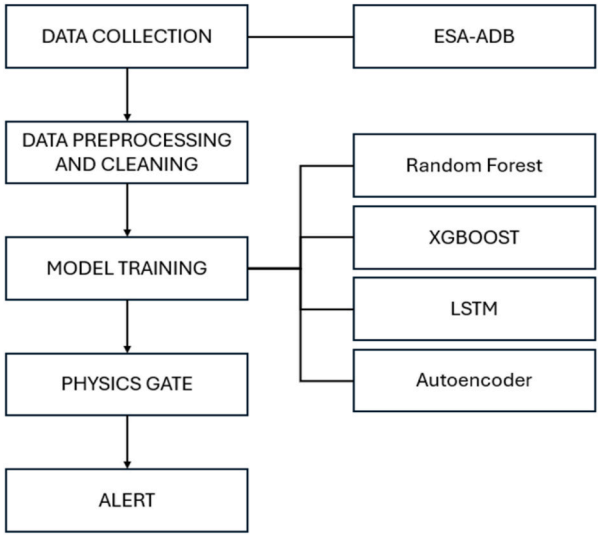

Fig. 2. Physics-informed anomaly detection pipeline.

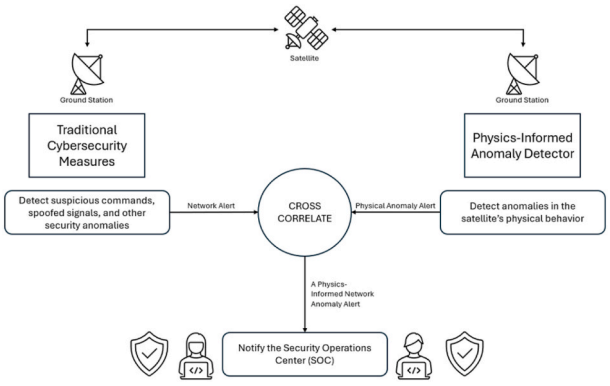

Fig. 3. Cross validation between traditional security measures and the proposed approach.

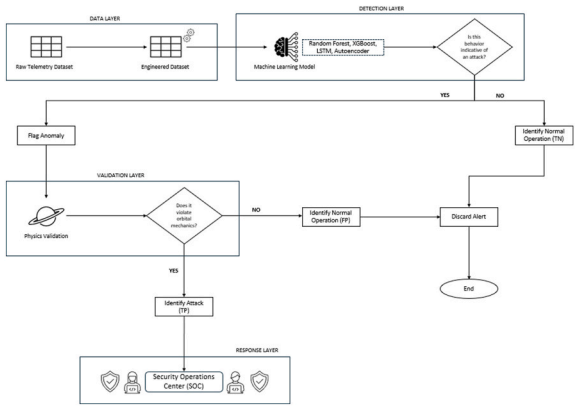

Fig. 4. Proposed hybrid system architecture to detect satellite orbital manipulation attacks.

the Detection Layer, the ML model (Random Forest, XGBoost, LSTM, or Autoencoder) analyzes the input data for anomalies. If the data is classified as "normal", this would mean that routine operations took place and would be considered as a true negative, concluding the workflow. If the model flags a potential attack, the case is passed to the Validation Layer. Here, the physics validation gate cross-checks the flagged anomaly against the fundamental laws of orbital mechanics using the instantaneous state vectors derived from the telemetry. If no physics violations are detected, the alert is classified as a false positive and discarded. Conversely, if a physics violation is confirmed, it is classified as a true positive and the Response Layer is triggered where the Security Operations Center (SOC) is immediately notified to initiate triage.

While prior efforts in physics-informed ML typically integrate physical laws into the training phase often via modified loss functions to regularize the learning process, this research introduces a novel active runtime gate. This distinction is critical since rather than merely helping the model "learn" physical patterns during training, the proposed architecture provides a deterministic audit of the model's output at runtime. This hybrid pipeline ensures that the final detection alert is not a "black box" statistical prediction, but a physically validated proof of state violation. By decoupling the statistical detection (ML layer) from the physical validation, a system that satisfies the high-integrity requirements of satellite mission control, where operators require deterministic evidence before taking defensive action is provided.

The novelty of this hybrid approach lies in its departure from standard Physics-Informed Machine Learning (PIML) paradigms. Traditional PIML methods typically integrate physical laws into the loss function during training to bias the model toward physically consistent solutions. While effective for data-starved scenarios, this "soft-constraint" approach still allows for "black-box" errors at runtime. In contrast, this research introduces a deterministic runtime gate that acts as a post-processing auditor. This solves a challenge unique to satellite operations – the need for explainable defensive action. In a high-stakes mission control environment, a statistical probability (e.g., ML) is insufficient to justify a satellite maneuver or lockout. However, a deterministic proof of energy conservation violation (i.e., a physics gate) provides the high-confidence forensic evidence required for operational decision-making. By decoupling detection from validation, the system remains robust even if the attacker attempts to maliciously trick the ML layer, as the physics gate remains an immutable barrier governed by orbital mechanics, not learned parameters.

To ensure clarity between the designed framework and the current implementation, the physics validation gate executed in this study focuses exclusively on the Conservation of Specific Orbital Energy (Vis-Viva Equation). The gate processes the scalar magnitudes of position and velocity from the telemetry over a ±7,200.00s window to calculate the energy discrepancy between consecutive states. A statistical alert is confirmed as a physical anomaly only if the energy discrepancy exceeds the adaptive tolerance of 51,000.00 J/kg, which accounts for natural LEO perturbations. Alerts falling within this tolerance are discarded as physically plausible noise, while the Tsiolkovsky and Angular Momentum gates remain reserved for future validation using full-vector and thrust-specific datasets.

A Data Collection

The ESA-ADB is a publicly available, large-scale, real-life dataset and framework from the ESA [33,34]. This dataset has anomalies curated and annotated, featuring 31.00 GB of data from three separate space missions [33,34]. It was curated to address the problems associated with inconsistent methods across the space industry for detecting and classifying satellite telemetry anomalies [33,34]. The orbital manipulation data in the ESA-ADB dataset is used in this research as a proxy for the kinematic footprint of a potential cyberattack, given the unavailability of publicly shareable, confirmed cyber-attack datasets. This allows for a reproducible and ethical methodology to validate the model's capabilities in detecting anomalous behavior.

B Data Preprocessing and Cleaning

During this stage, the raw telemetry data from Mission 01 was integrated and filtered to focus on telecommand-related anomalies. The preprocessing pipeline involved loading channel metadata (76 channels), anomaly labels (3589 anomalies across 200 events), and telecommand execution records (698 commands). A temporal filtering strategy was applied to retain only anomalies occurring within ±48.00 h of telecommand executions, reducing the dataset to 121,158 relevant telemetry records. The data was processed in chunks for memory efficiency, and anomaly windows were labeled into distinct phases, namely, 1) *pre-anomaly period*, 2) *active period*, and 3) *post-anomaly period*, resulting in a final dataset of 39,378 temporal snapshots. These preprocessing activities were efficiently performed using tools like Pandas and NumPy.

C Feature Engineering

The focus of the feature engineering phase was to create new features from raw data, remove irrelevant or harmful features, and optimize the feature set for model performance. As such, feature engineering was performed to create a comprehensive representation of telemetry behavior. From the available 76 channels, 27 channels were selected based on their anomaly involvement, effect size analysis, subsystem coverage, and physical unit diversity. The top 10 most affected channels during anomalies including channel_26 (showed increments in its values during anomalies), channel_41 (showed decrements in its values during

anomalies), and channel_28 (showed increments during anomalies) were prioritized. For each selected channel, three feature types were engineered, namely, 1) *raw telemetry values* representing direct sensor measurements, 2) *rate features* computed as first-order derivatives to capture the speed of change, and 3) *deviation features* calculated as the difference from a 12-point rolling mean baseline to identify unusual behavior patterns. This triple representation strategy resulted in 81 predictive features (27 channels $\times$ 3 feature types), providing models with multiple perspectives on telemetry dynamics for effective anomaly detection. These activities were carried out using tools and technologies like Pandas.

To ensure the validity of the performance metrics across all model families (Random Forest, XGBoost, LSTM, and Autoencoders), a strict causal boundary was maintained during feature engineering and preprocessing as further described below.

- Temporal Isolation: Rolling baselines, derivatives, and delta-based features were computed strictly within the temporal bounds of each data split. No "look-ahead" logic was employed and features at time t were derived solely from observations $\leq t$.
- Non-Recurrent Integrity: For non-recurrent tree models (e.g., Random Forest and XGBoost), rows were treated as independent observations only after causal feature engineering was completed, ensuring that the historical context of the orbit was preserved without peeking across the train/test split boundary.
- Per-Split Scaling: Global normalization parameters like mean and standard deviation for neural network-based models (e.g., LSTM and Autoencoders) were calculated exclusively on the training set and then applied to the validation and test sets. This prevents "distributional leakage" where the model could implicitly learn the range of the test data during training.
- Label Removal: To prevent label leakage, features directly derived from or semantically linked to the ground truth anomaly labels provided by ESA were systematically excluded. This included explicitly named metadata features as well as any diagnostic features provided in the raw dataset.

D Data Splitting and Validation

The focus of this phase was to create training, validation, and test sets that reflect real-world deployment, ensure temporal integrity for time series data as well as prevent data leakage between sets. Accordingly, temporal splitting was conducted to have a train:validation:test ratio of 7:1:2. For fairer comparison, this split was constant throughout all models. Additionally, stratified random splits were employed when timestamps were unavailable. As such, this phase leveraged tools and technologies like Pandas and Scikit-Learn.

E. Data Balancing and Sampling

The focus of this stage was to address class imbalance (if present) for better model training, create balanced datasets while preserving important patterns, and optimize for target metrics (priority was given to recall due to the nature of this research). The finalized dataset had an normal:anomaly ratio of 2:1. The dataset had a total number of 39,378 samples out of which 13,126 (33.33%) were anomaly samples whereas the remaining 26,252 (66.67%) were normal samples. Additionally, temporal preservation was maintained by sorting data chronologically before applying sequential splits. As such, for the above activities, tools and technologies like NumPy and Pandas were used.

F. Feature Scaling and Normalization

The focus of this phase was to normalize feature scales for algorithms which were sensitive to scaling, prepare data for neural networks and distance-based algorithms, and handle outliers and extreme values. Accordingly, algorithm-appropriate scaling, RobustScaler in particular, was used across all models to ensure consistency. As such, tools and technologies like Scikit-Learn and NumPy were used for this purpose.

G. Model Training

Experimentation for best model performance for problem resolution was done using four model families, namely, Random Forest, XGBoost, LSTM, and Autoencoder. Each model family had three model variants trained under them using a progressive refinement strategy, each better than its preceding variant.

1) Random Forest

Random Forest was selected for this research due to it being a gold standard for tabular data as it provides an established benchmark for structured datasets with mixed feature types. Random Forest models also provide robust default performance with minimal hyperparameter tuning required and reducing experimental bias. Additionally, Random Forest provides inherent anomaly detection as tree-based decisions naturally identify outliers through path isolation.

The baseline Random Forest model used all 81 features from the finalized dataset and no additional feature engineering was performed to ensure consistency across all models. The training strategy for this baseline Random Forest model used a default threshold of 0.50 and used default class weights.

The high recall focused Random Forest model focused on improving the recall of the baseline model. As such, several changes were introduced to the configuration of the baseline model and its hyperparameters. It also used heavy class weighting with a 50:1 penalty for false negatives. However, for fairer comparison, this model also used the same 81 features of the finalized dataset. The training strategy for this high recall focused Random Forest model used a slightly deeper max_depth value for more recall and made the min_samples_split and min_samples_leaf values slightly less restrictive than the baseline model. This high recall focused Random Forest model was trained on an optimal threshold of 0.46. Due to its smaller min_samples_leaf value, this model had more detailed leaf nodes compared to the baseline model.

The balanced precision-recall Random Forest model was optimized for an improved version of the high recall focused model. Similar to the baseline model and the high recall focused model, this model also used the same 81 features of the finalized dataset. This model also leveraged hyperparameter tuning and a balanced class weighting of 10:1. The training strategy for this optimized Random Forest model used values between those used for the baseline and high recall focused models for max_depth and min_samples_split. However, it used the same value as the baseline model for min_samples_leaf for stability. This optimized Random Forest model was trained on an optimal threshold of 0.43.

Table 1, Table 2, and Table 3 show a comparison between the preprocessing strategies, the model architectures and the thresholds of the three Random Forest models, respectively.

2) XGBoost

XGBoost was selected for this research due to its state-of-the-art tabular performance. XGBoost has a sophisticated regularization due to its built-in overfitting prevention capabilities which is crucial for

Table 1
Comparison of the preprocessing strategies of the random forest models.

Strategy	Baseline	High Recall	Balanced Precision-Recall
No. of Features	81	81	81
Data Split	7:1:2	7:1:2	7:1:2
Class Weight Strategy	Balanced	Heavy (50:1 Penalty)	Moderate (10:1 Ratio)

Table 2
Comparison of the model architectures of the random forest models.

Parameter	Baseline	High Recall	Balanced Precision-Recall
n_estimators	300	300	300
max_depth	20	25	22
min_samples_split	10	8	9
min_samples_leaf	5	4	5
max_features	'sqrt'	'sqrt'	'sqrt'
random_state	42	42	42
n_jobs	−1	−1	−1
verbose	1	1	1

Table 3
Comparison of the thresholds used to train each random forest model.

Model	Threshold
Baseline	0.50
High Recall	0.46
Balanced Precision-Recall	0.43

anomaly detection. It also has native support for sample weighting and class balancing techniques as well as optimized implementation, making it suitable and computationally efficient for large datasets. XGBoost also allows fine-tuning for specific recall/precision requirements, making it an ideal choice for this research study.

The baseline XGBoost model used standard preprocessing and leveraged all 81 features of the finalized dataset. It used a max_depth value of 6.00 for better capacity and calculated weights instead of manual weights based on the class ratio. It made use of reg_alpha and reg_lambda values to provide L1 and L2 regularization respectively and also used a min_child_weight of 3.00 to prevent overfitting. The training strategy for this baseline XGBoost model used a default threshold of 0.50.

The high recall focused XGBoost model was focused on improving the recall of the baseline model. As such, several changes were introduced to the baseline model. This model also used the same 81 features from the finalized dataset. The max_depth value used in this model was increased to capture more complex patterns whereas the min_child_weight was reduced to make this model less restrictive than the baseline XGBoost model. Compared to the baseline model, this model leveraged a heavy scale_pos_weight with a boost of 3.00-5.00x to prioritize the anomaly class. This high recall focused XGBoost model was trained on an optimal threshold of 0.30.

The balanced precision-recall XGBoost model also leveraged the same number of features as its counterparts but selected a max_depth value which was between the baseline and high recall focused XGBoost models and reverted to using the true balance value without any boosting for the scale_pos_weight value. The training strategy for this optimized XGBoost model used an optimal threshold of 0.43.

Table 4 and Table 5 show a comparison between the regularization strategies and the thresholds of the three XGBoost models, respectively.

3) LSTM

LSTM was selected for this research due to its temporal pattern recognition capabilities which are suitable for data with orbital periodicity. Due to its long-term dependencies, it can also capture orbital cycles and manipulation patterns over time. Its flexible architecture also makes it more adaptable to different sequence lengths and feature combinations.

To ensure zero data leakage and prevent sample adjacency bias, a strict chronological split was implemented. The dataset was ordered by the time index T before any partitioning. The training set contains data for $T < t_1$, the validation set for $t_1 < T < t_2$, and the test set for $T > t_2$. Feature scaling using RobustScaler was performed independently on

Table 4
Comparison of the regularization strategy of the XGBoost models.

Parameter	Baseline	High Recall	Balanced Precision-Recall
n_estimators	200	200	200
max_depth	6	7	6
learning _rate	0.10	0.10	0.10
scale_pos_weight	scale_pos_weight	scale_pos_weight	scale_pos_weight
reg_alpha	1.00	1.00	1.00
reg_lambda	5.00	5.00	5.00
subsample	0.80	0.80	0.80
colsample_bytree	0.80	0.80	0.80
colsample_bylevel	0.80	0.80	0.80
min_child_weight	3	2	3
objective	binary:logistic	binary:logistic	binary:logistic
eval_metric	aucpr	aucpr	aucpr
random_state	42	42	42
n_jobs	−1	−1	−1
verbosity	0	0	0

Table 5
Comparison of the thresholds used to train each XGBoost model.

Model	Threshold
Baseline	0.50
High Recall	0.30
Balanced Precision-Recall	0.43

each split to ensure that information from the test set (such as the global median or interquartile range) did not leak into the training process.

Furthermore, for the LSTM lookback window of 12 timesteps, windows were constructed within each split independently. Specifically, the first 12 timesteps of the test set were not used as a "target" to ensure that no hidden state information from the validation or training periods could propagate into the test evaluations. This temporal isolation ensures that the model is evaluated by its ability to forecast and detect anomalies in a truly "unseen" future state.

The baseline LSTM model used the same 81 features of the finalized dataset and used a standard architecture of 32 LSTM units in the first layer (L1) and 16 units in the second layer (L2). It used 12 timesteps and balanced class weights based on a normal:anomaly ratio of 2:1. The training strategy for this baseline LSTM model used a default threshold of 0.50.

The high recall focused LSTM model was focused on further tuning the baseline model to obtain an even better recall. As such, several changes were introduced to the baseline model. One such change was the use of a similar architecture that leveraged 32 LSTM units in the first layer (L1) and 16 units in the second layer (L2). This model also used 12 timesteps, although with a boosted class weight for anomalies of 1.20x. The training strategy for this high recall focused LSTM model used an optimal threshold of 0.55.

The balanced precision-recall model was focused on improving the precision-recall tradeoff of the high recall focused model. As such, it was similar to the high recall focused model in several aspects. For example, it used the same 81 features of the finalized dataset and the same number of timesteps (12). Similarly, this model also used 32 LSTM units in the first layer (L1) and 16 units in the second layer (L2), just like the other two model variants and also used a balanced weight class of 2:1 normal:anomaly ratio, similar to the baseline LSTM model. The training strategy for this balanced precision-recall LSTM model used an optimal threshold of 0.85. This model was trained with a high dropout rate on the first layer for stronger regularization.

Table 6, Table 7, and Table 8 show a comparison between the feature dimensionality strategy, the model complexity evolution, and the thresholds of the three LSTM models, respectively.

Table 6
Comparison of the feature dimensionality strategy of the LSTM models.

Dimensions	Baseline	High Recall	Balanced Precision-Recall
No. of Layers	32	16	24
No. of Timesteps	12	12	12
Class Weight	2:1	2.40:1	2:1
No. of Epochs	50	50	100

Table 7
Model complexity evolution of the LSTM models.

Parameter	Baseline	High Recall	Balanced Precision-Recall
LSTM Units (L1)	32	32	32
LSTM Units (L2)	16	16	16
Dense Layers	2	2	3
Dropout Rate (L1)	0.30	0.20	0.40
Dropout Rate (L2)	0.30	0.20	0.40

Table 8
Comparison of the thresholds used to train each LSTM model.

Model	Threshold
Baseline	0.50
High Recall	0.55
Balanced Precision-Recall	0.85

4) Autoencoders

Autoencoders were selected for this research due to their true anomaly detection paradigm as it learns normal patterns and flags deviations, making it theoretically ideal for the detection of orbital manipulation attacks. They also use an unsupervised approach that provides a practical advantage in its use. Autoencoders also have reconstruction-based detection capabilities, making it a more natural fit for identifying corrupted orbital parameters. It also provided representation learning capabilities as it was capable of discovering compact representations of normal orbital states.

The baseline Autoencoder model used the same 81 features of the finalized dataset that was used by the other models and is a simple dense Autoencoder with a fully connected feedforward network. It comprised of an encoder with two dense layers that compressed the input to a small vector using an encoding_dim value of 32. The decoder consisted of symmetric dense layers that reconstructed the original input. This model used L2 regularization, batch normalization, and dropout values to provide stabilization and prevent overfitting. Furthermore, the reconstruction accuracy of this model was measured using Mean Squared Error (MSE). The training strategy for this baseline Autoencoder model used a default threshold of 0.17 and used an Adam optimizer value of 0.0010.

The Variational Autoencoder (VAE) model was focused on extending the baseline autoencoder model into a probabilistic model by using the same 81 features of the finalized dataset. This model learnt latent distributions instead of fixed encoding via values like z_mean (mean of latent distribution) and z_log_var (log variance). To introduce stochasticity, it uses reparameterization by combining the lambda layer with sampling. However, it leveraged the encoder and the decoder as separate models and included a custom loss layer that combines reconstruction loss measured via MSE and Kullback-Leibler (KL) divergence (D_{KL}) controlled by the weighting factor β. The training strategy for this VAE model used an optimal threshold of 0.21 and used an Adam optimizer value of 0.0005.

$$Loss = MSE + \beta . D_{KL} \quad (1)$$

The high recall focused Autoencoder was focused on taking a deeper, more expressive approach focused on robustness and better generalization for increased tolerance to outliers and noise compared to the other Autoencoder models. This model also utilized the same 81 features of the finalized dataset and employed a symmetric architecture with more layers for high learning capacity. This model used Huber loss function as it was more robust to outliers. The training strategy for this high recall focused Autoencoder model used an optimal threshold of 0.30 and used an Adam optimizer value of 0.0010.

Table 9 and Table 10 show the evolution of the feature engineering strategy and the thresholds of the three Autoencoder models, respectively.

The classification thresholds for all models were determined via a systematic search procedure. For the Random Forest and XGBoost model variants, the precision_recall_curve was utilized on the validation set to identify the lowest threshold satisfying an operational recall floor (≥90.00%) whereas for the LSTM and Autoencoder model variants, the thresholds were selected through a discrete sweep across the [0.10, 0.90] range with step sizes of 0.05 absolute increments and 1.00 percentile increments respectively, optimized to maximize the F1-Score subject to a high-integrity recall constraint (≥95.00%).

H. Physics Validation

1) Orbital Kinematics and Physical Principles

A fundamental challenge in space cybersecurity research is the absolute lack of publicly available telemetry from confirmed satellite cybersecurity incidents. This research addresses this gap by utilizing command-related orbital anomalies from the ESA-ADB dataset as a kinematic proxy. The scientific rationale relies on the principle of physical equifinality, where from a flight dynamics perspective, an unauthorized change in orbital state, whether induced by sensor spoofing or malicious telecommand injection, is kinematically indistinguishable from a legitimate but anomalous state change, provided their thrust profiles are similar.

By focusing on the "kinematic footprint" of the manipulation, a system capable of validating the physical legitimacy of the reported orbital state is inherently capable of detecting a cyber-induced manipulation attempt, regardless of the attacker's specific exploit vector.

A novel component introduced in this research is the proposed multi-gate physics architecture. Physics validation acts as a crucial post-processing filter by confirming predictions that are mathematically impossible (e.g., gross energy discrepancies), while filtering out those that are physically plausible (statistical noise). As such, having a multi-gate physics architecture means the detected anomalies represent severe physical impossibilities that cannot be explained by known orbital mechanics, thereby dramatically increasing the anomaly detection system's reliability.

While this study focuses on "loud" footprints (major deviations in orbital energy), a "stealth boundary" defined by the calibrated physics threshold (51,000.00 J/kg) is acknowledged as follows.

- Loud Attacks (>51,000.00 J/kg): Identified deterministically by the physics gate as physical impossibilities.
- Stealth Attacks (≤51,000.00 J/kg): Scenarios where an adversary attempts to "hide" a maneuver within natural perturbation noise. Here, the system relies on the ML layer's temporal sensitivity to

Table 9
Evolution of the autoencoder feature engineering strategy.

Feature Data	Baseline	VAE	High Recall
Input Features	81	81	81
Latent Space	Deterministic	Probabilistic	Deterministic
Loss Function	MSE	MSE + D_{KL}	Huber
Dropout	0.10	0.20	0.10
Output Models	One Model	Three Models	One Model

Table 10
Comparison of the thresholds used to train each autoencoder model.

Model	Threshold
Baseline	0.17
VAE	0.21
High Recall	0.30

identify the cumulative statistical "drift" that differentiates a malicious maneuver from expected orbital decay.

This dual-layer approach ensures that while the physics gate provides a high-confidence "hard floor" for data integrity, the ML layer remains the primary defense against subtle, unauthorized behaviors.

The most critical feature of the physics gate is the instantaneous check over a tiny time window. It calculates the difference in orbital state immediately before and immediately after the anomaly event. In this research, this time window is defined as ±7200.00s. The primary driver for this interval is the underlying sampling rate of the telemetry stream. The system operates with a median cadence of approximately 7200.00s. This means that when the anomaly timestamp is detected, the nearest valid preceding and succeeding data points are by definition, 7200.00s apart (or multiples thereof). Thus, the interval is technically forced by the data availability, as a smaller window would not contain two distinct data points for comparison.

Furthermore, for a typical LEO satellite, attitude changes, thruster burns, and resulting orbital adjustments (due to the execution of telecommands) are often scheduled and executed over intervals corresponding to ground station passes or full orbit periods, which naturally align with or exceed the 7200.00s window [35]. As such, using the ±7200.00s interval ensures that the validation captures the entire, accumulated kinematic effect of any command that was issued around the time of the anomaly.

The hybrid architecture proposed in this research addresses two distinct classes of orbital manipulation threats. The physics validation gate is designed to detect kinematically inconsistent anomalies, such as sensor spoofing or data injection, where the reported state violates conservation laws (e.g., a sudden orbital "jump" without a corresponding energy change). Conversely, the ML component is responsible for identifying physically plausible but unauthorized maneuvers.

For example, a "stealthier" attack might involve a small, unauthorized thruster burn that stays within the calibrated 51,000.00 J/kg energy tolerance. While such a maneuver is physically possible and would not trigger the physics gate, it represents an unexpected deviation from the learned "normal" operational pattern, which the ML layer is tuned to detect as a statistical anomaly. By maintaining this distinction, the system ensures that gross physical inconsistencies are discarded as high-confidence threats, while subtle, unauthorized behaviors are flagged for further operator investigation.

This component mainly considers Newton's Law of Universal Gravitation, Newton's Second Law for Circular Motion, Law of Conservation of Specific Orbital Energy, The Tsiolkovsky Rocket Equation Principle, and Law of Conservation of Angular Momentum.

Newton's Law of Universal Gravitation states that every particle in the universe attracts every other particle with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between their centers [36]. This is the source of the orbital motion. It defines the potential energy component of the orbit and sets the exact acceleration experienced by the satellite. It is required to calculate the total specific energy and to determine the necessary escape velocity, ensuring the satellite is on a bounded orbit, not just moving arbitrarily. It is mathematically expressed as follows:

$$F_g = G\frac{Mm}{r^2} \tag{2}$$

- F_g is the gravitational force between the two objects [36].
- G is the gravitational constant, a value experimentally determined to be approximately $6.674 \times 10^{-11}\ Nm^2/kg^2$ [36].
- M and m are the masses of the two objects [36].
- r is the distance between the centers of the two masses [36].

Newton's Second Law for Circular Motion states that the net force acting on an object moving in a circle is equal to its mass times its centripetal acceleration [36]. This net force is directed toward the center of the circle and is called the centripetal force [36]. This law can therefore be used to calculate the Keplerian velocity, serving as a baseline check for the internal consistency of the reported radius and velocity values. It is mathematically expressed as follows:

$$F_c = ma_c \tag{3}$$

- F_c is the centripetal force (the net force causing the circular motion) [36].
- m is the mass of the object [36].
- a_c is the centripetal acceleration [36].

Using centripetal acceleration which is mathematically expressed as $a_c = \frac{v^2}{r}$, the above equation could be expressed as follows:

$$F_c = m\frac{v^2}{r} \tag{4}$$

- v is the object's tangential speed [36].
- r is the radius of the circle [36].

The relationship between a satellite that circles the earth close to its orbit can be expressed as $F_c = F_g$ [36]. Therefore, the gravitational orbital velocity can be expressed as follows:

$$\begin{aligned} m\frac{v^2}{r} &= G\frac{Mm}{r^2} \\ v^2 &= \frac{GM}{r} \\ v &= \sqrt{\frac{GM}{r}} \end{aligned} \tag{5}$$

The Tsiolkovsky Rocket Equation determines the maximum possible change in velocity that a spacecraft can achieve from a particular propulsive maneuver [37]. This is the single most important metric for sizing a rocket, as the change in velocity budget dictates how much mass can be moved from one orbit to another [37]. As such, this can be considered as the physical actuation limit as it links the change in velocity to the cause. It ensures that any detected change in velocity is less than the satellite's maximum physical capability. As such, violations of this equation flag events that are impossible to generate by the spacecraft itself, regardless of power or fuel constraints. It can be mathematically expressed as follows:

$$\Delta v = v_e \ln\left(\frac{m_0}{m_f}\right) \tag{6}$$

- Δv refers to the change in velocity [37].
- v_e is to the effective exhaustive velocity [37].
- m_0 is the total initial mass [37].
- m_f refers to the final mass [37].

For the purpose of this gate, the Δv threshold is set to 250.00 m/s, which is justified as it exceeds the entire annual station-keeping budget for typical LEO missions (50.00 m/s) and would violate propellant mass fraction constraints for the auxiliary thrusters used in the ESA missions [30]. This value ensures that any reported instantaneous change in velocity exceeding this limit flags a hard physical impossibility.

The Law of Conservation of Specific Orbital Energy states that in a

closed, isolated system where the only force acting on the orbiting body is gravity (a conservative force), the specific mechanical energy of the orbiting body remains constant at every point along its trajectory [38]. This is the primary integrity check for the orbit's magnitude. Energy is the scalar value that determines the size and shape of the orbit (e.g., circular, elliptical, hyperbolic, etc.). A sudden, massive, and uncompensated change in energy indicates a violation of energy conservation that cannot be achieved physically. It is the most robust test for data corruption. The physics validation gate of this research uses the Vis-Viva Equation, which is a direct form of the Law of Conservation of Specific Orbital Energy [38]. The Vis-Viva gate serves as the primary deterministic filter for identifying kinematically inconsistent orbital state changes. This module utilizes the Law of Conservation of Specific Orbital Energy to verify if the reported change in a satellite's state is kinematically consistent with a two-body gravitational model, adjusted for LEO perturbations. The gate ingests telemetry pairs consisting of instantaneous scalar velocity (v) and radial distance (r) from the Earth's center at two time-steps.

The specific mechanical energy of the satellite is calculated at two discrete points: immediately before (t_{pre}) and after (t_{post}) a detected anomaly. The instantaneous energy is defined as:

$$\in = \frac{v^2}{2} - \frac{\mu}{r} \tag{7}$$

- $\in$ refers to the specific orbital energy [38].
- v is the instantaneous orbital speed of the body [38].
- μ is the Standard Gravitational Parameter and can be expressed using (2) as $\mu = GM$ [38].
- r is the instantaneous distance between the body and the central mass [38].

An anomaly is confirmed as a physical violation if the absolute change in specific energy exceeds the calibrated threshold which can be expressed as:

$$\left|\in_{post} - \in_{pre}\right| > r \tag{8}$$

Events where $|\Delta \in| \leq r$ are classified as physically plausible (e.g., natural perturbations or valid maneuvers) and are filtered out as statistical noise.

The gate utilizes a fixed temporal window of $\pm$7200.00s around the anomaly timestamp. This interval is forced by the telemetry cadence (median 7200.00s), ensuring the inclusion of exactly two distinct data points for comparison. This window is sufficient to capture the full kinematic effect of typical LEO maneuvers or command executions.

The tolerance of 51,000.00 J/kg is characterized as mission-adaptive because it was scientifically calibrated using the Ansys STK ground truth "ephemeris" dataset. This reference data represents ideal physical behavior free from sensor error. To establish r, a high-fidelity High-Precision Orbit Propagator (HPOP) was configured with the following parameters:

- Orbit Profile: Sun-Synchronous Orbit (SSO) with a Semi-Major Axis (SMA) of 7071.00 km (approximately 700 km altitude) and an inclination of 98.20°.
- Force Models: Earth Geopotential (J_2 through J_4 harmonics) and Atmospheric Drag using the MSISE 1990 model.
- Derivation: By calculating the energy drift of this perturbed "real-world" simulation over the 7200.00s window, the maximum non-Keplerian residual was identified. The value of 51,000.00 J/kg was selected to absorb these natural variations (specifically the $\approx$50,000.00 J/kg drift observed in LEO). This ensures high sensitivity to malicious "jumps" in state while maintaining absolute robustness against natural orbital decay or Earth oblateness effects, as characterized by Vallado [39].

The Law of Conservation of Angular Momentum states that the total angular momentum of an isolated system remains constant in both magnitude and direction if and only if the net external torque acting on the system is zero [36]. This is the primary integrity check for the orientation of the orbit in this physics validation gate. Angular momentum defines the orbital plane [36]. Since angular momentum is a vector, its conservation ensures both the radius and velocity vectors are correctly oriented [36]. A sudden change in the direction without sufficient thrust is a hard physics violation, indicating an unphysical change in the satellite's orientation telemetry. This can be derived from Newton's Second Law for Circular Motion from (3) above and can be expressed as:

$$T_{net} = \frac{dL}{dt} \tag{9}$$

- T_{net} refers to the net external torque [36].
- L is the total angular momentum [36].
- t refers to the time [36].

The multi-gate physics architecture of this research makes the following assumptions.

- The satellite's motion is primarily governed by the gravitational force of a single, point-mass earth which is a perfect sphere. As such, it ignores all other forces and gravitational influences. However, to account for the neglected forces, an adaptive tolerance of 51,000.00 J/kg, calibrated against the generated ground truth data, is explicitly added causing the actual energy state to deviate from the perfect two-body prediction over the 7200.00s window [39].
- Any significant change in velocity is treated as an instantaneous event occurring between the two telemetry states.
- A small, fixed allowance is sufficient to cover sensor inaccuracies and measurement noise.
- The position and velocity data points recorded at a single timestamp are kinematically synchronized and represent the true state of the satellite at that precise moment.
- The relationship between the synthetic proxy features (radius and velocity) and their true physical values is linear and constant, as determined by the initial standard deviation calibration against the STK ground truth data.

To maintain technical transparency, this research distinguishes between validated and designed components. The Vis-Viva gate is implemented and empirically validated using the current dataset. The Tsiolkovsky gate and the Angular Momentum gate are designed for future validation; they are architecturally integrated but remain inactive in the current empirical results due to the specific limitations of proxy telemetry and the lack of propellant-specific attack data.

2) Implementation and Validation Scope

The proposed system incorporates a multi-gate physics validation framework designed to enforce the laws of orbital mechanics as a final integrity check. To ensure technical transparency, the components of this framework are categorized by their implementation status in this study as.

- Implemented and Empirically Validated (Vis-Viva Gate): The Vis-Viva gate is the primary operational component of this research. It utilizes the Law of Conservation of Specific Orbital Energy to identify gross kinematic inconsistencies in the reported orbital state. This gate was fully validated using the provided dataset and forms the basis for the precision improvements reported.
- Designed for Future Validation (Tsiolkovsky and Angular Momentum Gates): While architecturally integrated into the system's

logic flow, the following gates remain in a design state due to current data constraints:

- o Tsiolkovsky Gate: Designed to detect propulsion-limit violations; requires thrust-specific telemetry not present in the current proxy dataset.
- o Angular Momentum Gate: Designed to detect orbital plane manipulation; requires full 3D cartesian vectors for cross-product calculation, which exceeds the scope of the current scalar proxy mode.

3) Reference Simulation and Ground Truth Generation

Ansys Systems Tool Kit (STK), formerly Satellite Tool Kit, is a sophisticated digital mission engineering software used in the aerospace and defense industries to model, simulate, and analyze complex systems (e.g., satellites, aircraft, missiles, etc.) in a realistic, time-dynamic, and physics-based 3D environment and is used by major aerospace companies including NASA, ESA, Boeing, etc. [40].

For the purpose of the multi-gate physics architecture, a dataset was generated by the Ansys STK software to represent highly accurate, physics-based, simulated references for a mission. This data provides the ideal, known physical behavior of a satellite, free from real-world noise, sensor error, and telemetry corruption. In the multi-gate physics architecture, the STK ground truth data is used to ensure the physics gate operates on a correct scale, even when provided with abstract or synthetic telemetry data.

The STK generated dataset is a time-series record of the simulated satellite's precise state vectors, also known as "ephemeris". It includes several key fields such as.

- Time: The precise timestamp for each data point.
- Cartesian Position (x, y, z): The satellite's location in space, relative to the center of the earth.
- Cartesian Velocity (V_x, V_y, V_z): The instantaneous speed of the satellite and direction in space.
- Classical Elements (Semi-Major Axis (SMA), Eccentricity (Ecc), Inclination (Inc)): Orbital shape and orientation parameters.

This data was used to calibrate the proxy telemetry and to establish the baseline of normal, non-anomalous physical behavior against which the physics validation gate tests the highly corrupted data. Fig. 5 depicts the workflow of the physics validation gate. As shown in the diagram, the Vis-Viva gate is fully validated whereas the Tsiolkovsky and Angular Momentum modules are designed architectural components for future verification.

I. Model Evaluation

This model leverages several performance evaluation metrics like accuracy, precision, recall, and F1-Score, as well as several diagnostic metrics such as True Positive Rate (TPR), True Negative Rate (TNR), FPR, and FNR.

In the context of satellite security, a false negative or a missed real attack could cost millions of dollars and would have drastic security implications. As such, the primary metric of this research was prioritized to be the recall. If the recall is maximized but the precision is too low, there is going to be an overflow of false positives. In satellite operations, too many false positives could cause alert fatigue for the operators and/or automated mitigation systems. As such, precision was prioritized next to keep the alerts manageable and meaningful. Table 11 shows the target value for each metric that was set for this research.

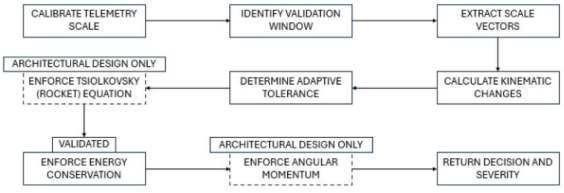

Fig. 5. Workflow of the physics validation gate.

The initial performance targets for accuracy, precision, and F1-Score were established at a baseline of >20.00% to serve as a "minimum viability floor" for this research. These targets were specifically calibrated during the earliest stages of the study when the dataset was highly imbalanced, and initial non-temporal models struggled to reach even 10.00% precision due to the high signal-to-noise ratio in operational telemetry. Setting these thresholds allowed for a quantitative assessment of whether ML could provide a meaningful "lift" over baseline statistical noise. While the final LSTM-based results significantly exceeded these preliminary benchmarks, the 20.00% target remains a critical historical metric for the project, representing the transition from non-viable detection to statistically significant signal identification.

5. Results and discussion

A Model Performance Results

1) Random Forest Models

Table 12 shows a summary of the results of the performance and diagnostic metrics for the Random Forest models. When considering the individual performance of the Random Forest models, the recall was the highest in the high recall focused Random Forest model (97.58%) while the lowest was for the balanced precision-recall Random Forest model (57.95%). In terms of precision, the highest value was observed for the baseline model (33.92%) while the lowest was observed once again for the balanced precision-recall model. The 95% Confidence Interval (CI) for the high recall focused model was found to be [33.50%, 33.81%] for the precision and [96.94%, 98.15%] for the recall. In terms of security, the high recall focused model performed the best as it missed only about 2.42% of the real attacks. However, in terms of operational efficiency, it generated about 97.40% of false positives. In a real-world satellite control center, an FPR this high would quickly lead to alert fatigue, rendering the system unusable and potentially masking genuine threats. This operational overhead is a critical limitation of the pure ML approach. When considering the balanced precision-recall model, it had a recall of 57.95%, thereby missing 44.06% of real attacks, which can be considered catastrophic. However, in terms of precision, it had a value of 33.37%, thereby generating only 58.35% of false positives, which is considerably lesser compared to the high recall focused Random Forest model. Nevertheless, as recall was the primary metric for this research, the high recall focused model can be considered as the best Random Forest model that was tuned.

Fig. 6 shows the Precision-Recall (PR) curves for the high recall focused Random Forest model. This model demonstrates substantial improvement in attack detection capability, achieving excellent recall values ($R_{validation} = 0.97, R_{test} = 0.98$) at optimal thresholds. The PR curve indicates that although fluctuations in precision are observed in the beginning, it stabilizes with the increase in recall, once again around 0.30 ($P_{validation} = 0.125, P_{test} = 0.110$).

Table 11
Target values for the proposed performance and diagnostic metrics.

Metric	Target Value
Accuracy	> 20.00%
Precision	> 20.00%
Recall	> 85.00%
F1-Score	> 25.00%
TPR	> 85.00%
TNR	> 40.00%
FPR	< 80.00%
FNR	< 10.00%

Table 12
Model evaluation for random forest models.

Metric	Baseline	High Recall	Balanced Precision-Recall
Accuracy	43.49%	34.53%	46.97%
Precision	33.92%	33.65%	33.37%
Recall	71.82%	97.58%	57.95%
F1-Score	46.07%	50.05%	42.35%
TPR	71.82%	97.58%	57.95%
TNR	29.15%	2.60%	41.62%
FPR	70.85%	97.40%	58.38%
FNR	28.18%	2.42%	44.06%

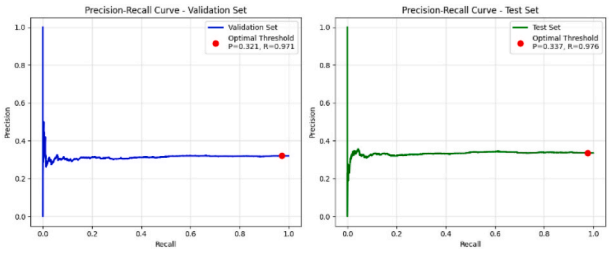

Fig. 6. PR curves for the random forest high recall focused model with a precision CI [33.50%, 33.81%] and recall CI [96.94%, 98.15%].

As shown in Fig. 7, the high recall focused model performed exceptionally well with a recall of 97.58%, thereby detecting 2583 true orbital manipulation attacks while missing only 64 true events. However, compared to the baseline model, this model generated an increased number of false positives due to its slightly lesser precision of 33.65%.

2) XGBoost Models

Table 13 shows a summary of the results of the performance and diagnostic metrics for XGBoost models. When considering the individual performance of the XGBoost models, it can be seen that the model with the best recall (95.81%) was the high recall focused model, thereby missing only 4.19% of real attacks. However, this model had a precision of only 33.60%, thereby resulting in 95.87% false positives. The 95% CI for the high recall focused model was found to be [33.37%, 33.83%] for the precision and [95.01%, 96.56%] for the recall. Similar to the Random Forest models, this extremely high FPR presents an unacceptable operational risk. Operators would be forced to investigate an overwhelming volume of non-threat alerts, draining resources and significantly lowering the trustworthiness of the security system. In terms of precision, the balanced precision-recall performed with a

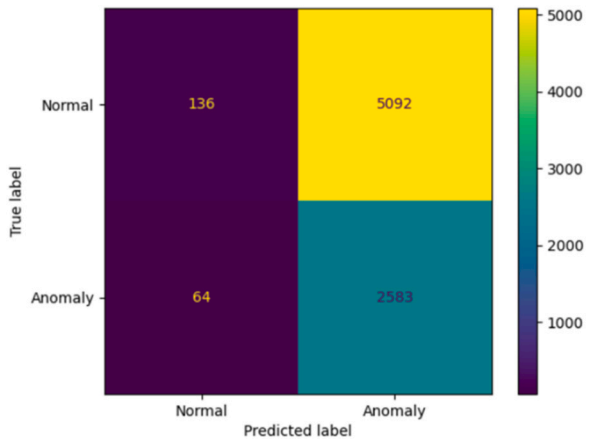

Fig. 7. Confusion matrix for the test set of the random forest high recall focused model.

Table 13
Model evaluation for XGBoost models.

Metric	Baseline	High Recall	Balanced Precision-Recall
Accuracy	54.72%	34.95%	46.30%
Precision	32.31%	33.60%	33.10%
Recall	31.70%	95.81%	58.52%
F1-Score	31.25%	49.75%	42.28%
TPR	31.70%	95.81%	58.52%
TNR	66.37%	4.13%	40.11%
FPR	33.63%	95.87%	59.89%
FNR	68.30%	4.19%	41.48%

precision of 33.10% while maintaining a recall of 58.52% that resulted in missing 41.48% of real attacks while generating 59.89% of false positives. The baseline model was the most underperforming model of the three XGBoost models with a recall of 31.70% and a precision of 32.31%. As such, in terms of security, the high recall focused XGBoost model can be recommended although operational efficiency may be at stake due to the large number of false positives generated.

Fig. 8 shows the PR curves for the high recall focused XGBoost model. This model achieves excellent attack detection capability with high recall values ($R_{validation} = 0.97, R_{test} = 0.96$), demonstrating effective optimization for the primary metric. The PR curves show stable, flat profiles indicating consistent precision across recall ranges. This model also exhibits positive transfer from validation to test, with improved performance on precision ($P_{validation} = 0.32$ *to* $P_{test} = 0.34$), suggesting robust generalization. Compared to the Random Forest high recall focused model, XGBoost maintains a slightly lower recall performance while achieving an almost similar precision, indicating comparable effectiveness for anomaly detection.

As shown in Fig. 9, the high recall focused XGBoost model performed exceptionally well with a recall of 95.81%, thereby detecting 2536 true orbital manipulation attacks while missing only 111 true events in the test set. However, it also has to be noted that this model generated a large number of false positives due to its low precision of 33.60%.

3) LSTM Models

Table 14 shows a summary of the results of the performance and diagnostic metrics of the LSTM models. When considering the individual performance of the LSTM models, it can be seen that a recall of 95.22% was observed for the baseline LSTM model. However, when comparing the other metrics for the same model, it can be seen that this model was well balanced overall. As such, in addition to a recall of 95.22%, it also had an FPR of 9.35%. As such, this model can be considered to be of great use in a real-world setting. Similarly, when comparing the remaining LSTM models, it can be seen that the balanced precision-recall LSTM model had a recall of 95.64% and a precision of 90.88%. Furthermore, it had an FPR of 4.84% and it only missed 4.36% of real attacks. This is a significant improvement in terms of operational efficiency from the high recall focused LSTM model which had a recall of 95.71%, resulting in a FNR of 4.29% and FPR of 5.72%. The 95% CI for the high recall focused model was found to be [89.88%, 91.87%] for the

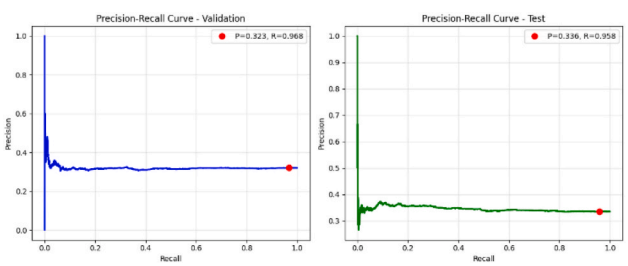

Fig. 8. PR curves for the XGBoost high recall focused model with a precision CI [33.37%, 33.83%] and recall CI [95.01%, 96.56%].

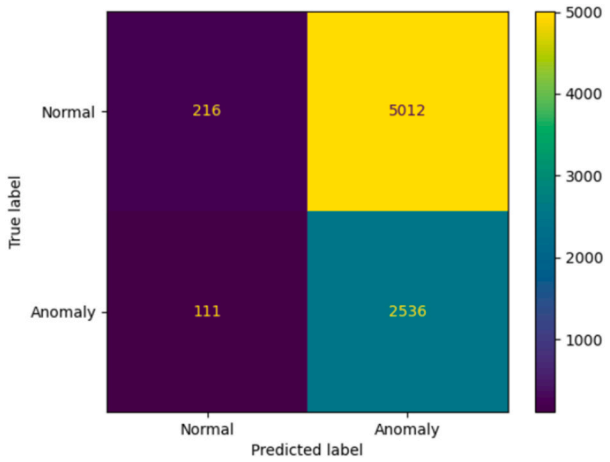

Fig. 9. Confusion matrix for the test set of the XGBoost high recall focused.

Table 14
Model evaluation for LSTM models.

Metric	Baseline	High Recall	Balanced Precision-Recall
Accuracy	92.18%	94.76%	95.32%
Precision	83.69%	89.40%	90.88%
Recall	95.22%	95.71%	95.64%
F1-Score	89.09%	92.45%	93.20%
TPR	95.22%	95.71%	95.64%
TNR	90.65%	94.28%	95.16%
FPR	9.35%	5.72%	4.84%
FNR	4.78%	4.29%	4.36%

precision and [94.86%, 96.39%] for the recall. In the security sense, although the balanced precision-recall LSTM model performed slightly lesser than the high recall focused model, from an operational perspective, this model performed the best, thereby avoiding catastrophic outcomes and operational overhead observed in the other model families.

Fig. 10 shows the PR curves for the balanced precision-recall model of LSTM. This model achieves excellent recall performance ($R_{validation} = 0.94, R_{test} = 0.96$) showcasing great detection capabilities while maintaining comparable precision ($P_{validation} = 0.90, P_{test} = 0.91$). The PR curves display flat, stable profiles indicating consistent performance across threshold ranges. The model demonstrates positive generalization with improved test performance ($R + 2.00\%, P + 1.00\%$), suggesting effective regularization without any aggressive performance degradation. This represents the optimal LSTM configuration, combining near-perfect attack detection with stable precision and robust generalization.

When considering the balanced precision-recall LSTM model that was trained to achieve a better precision-recall tradeoff from the high recall focused model, it can be seen that this model performed better compared to the baseline model with a recall of 95.64% with only a slight reduction in its value when compared to the high recall focused model. This meant that as shown in Fig. 11, this model correctly detected 2521 attacks in the test set while the number of false positives generated from this model was much less than the number of false positives from the other two variants, thereby showcasing its strengths at anomaly detection in satellite telemetry data, with reduced operational overhead.

4) Autoencoders

Table 15 shows a summary of the results of the performance and diagnostic metrics of the Autoencoder models. When considering the individual performance of the Autoencoder models, it can be seen that the model with the best recall was the high recall focused Autoencoder model with a recall of 90.29%. This model had a precision of 33.62% and an FPR of 90.26%, while missing only 9.71% of real attacks. The 95% CI for this model was found to be [33.24%, 34.00%] for the precision and [89.22%, 91.48%] for the recall. However, when comparing its other variants, it can be seen that the baseline Autoencoder model had a recall of only 48.85% and the VAE variant had a recall of only 36.76%. From an operational perspective, even the best-performing Autoencoder variant is fundamentally unsuitable for this high-stakes application. This poor performance, particularly the high FPR, is not a matter of hyperparameter tuning or minor architectural choice, but is attributed to the core misalignment of Autoencoders with the supervised problem structure.

- Misalignment with Task Structure: Autoencoders are primarily unsupervised tools designed for outlier detection via reconstruction error. Because this research utilizes an explicitly labeled dataset for binary classification, supervised architectures (RF, XGBoost, LSTM) naturally outperform the Autoencoder's attempt to implicitly learn a "normal" boundary.
- Impact of Class Balance: Autoencoders rely on a "heavy" majority class to learn a clean baseline. The 2:1 (normal:anomaly) ratio in this filtered dataset meant the models were forced to "learn" anomalous patterns as part of the normal state reconstruction, leading to a blurred decision boundary and the observed high FPRs (up to 90.26%).
- Lack of Temporal Context: Unlike the LSTM model family, which uses hidden states to link sequential data points, the dense Autoencoders used in this study treated each telemetry snapshot as an independent instance. Consequently, they failed to capture the time-linked kinematic evolution necessary to distinguish a smooth orbital maneuver from a sudden manipulation attack.

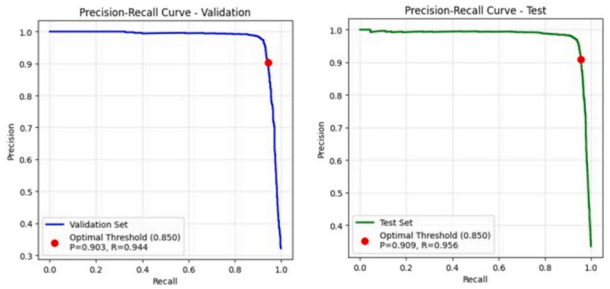

Fig. 10. PR curves for the LSTM balanced precision-recall model with a precision CI [89.88%, 91.87%] and a recall CI [94.86%, 96.39%].

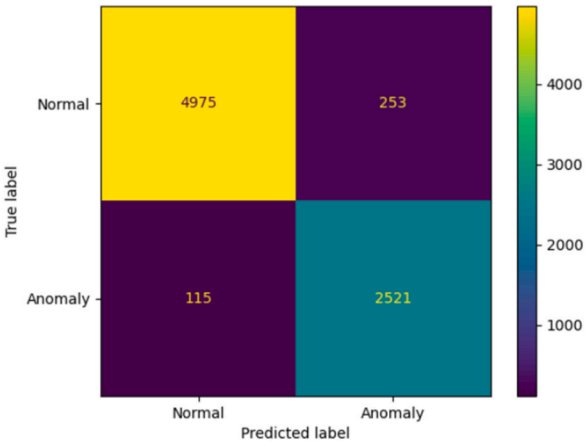

Fig. 11. Confusion matrix for the Test Set of the LSTM Balanced Precision-Recall Model.

Table 15
Model evaluation for autoencoder models.

Metric	Baseline	VAE	High Recall
Accuracy	50.78%	53.63%	36.81%
Precision	33.89%	32.97%	33.62%
Recall	48.85%	36.76%	90.29%
F1-Score	40.02%	34.76%	49.00%
TPR	48.85%	36.76%	90.29%
TNR	51.76%	62.17%	9.74%
FPR	48.24%	37.83%	90.26%
FNR	51.15%	63.24%	9.71%

Fig. 12 shows the PR curves for the high recall focused Autoencoder model. This model demonstrates the best performance within the Autoencoder family, achieving substantially improved recall ($R_{validation} = 0.91, R_{test} = 0.90$) while maintaining reasonable precision ($P_{validation} = 0.32, P_{test} = 0.34$). The PR curves show enhanced discriminative capability with steeper initial drops, indicating better threshold sensitivity. This configuration finally approaches competitive performance levels, though still substantially below LSTM and tree-based methods for orbital manipulation attack detection.

When considering the high recall focused Autoencoder model that was trained to achieve a better recall than the two other Autoencoder models, it can be seen that this model performed the best with a recall of 90.29%. This meant that the high recall focused model only missed 257 true attacks from the test set but detected 2390 true attacks successfully. However, as shown in Fig. 13, the number of false positives generated from this model was higher compared to the other two Autoencoder models.

A deeper architectural comparison reveals that the Autoencoders underperformed primarily because they were implemented as static, fully connected (dense) networks. These point-in-time models suffered from an objective mismatch; while the goal was to detect anomalous transitions, the models were optimized solely to minimize reconstruction error for isolated snapshots. Consequently, the Autoencoder could reconstruct an anomalous radius or velocity value with high fidelity if that specific value was statistically prevalent in the training set, even if the transition leading to that value was kinematically impossible.

The implementation of more advanced temporal architectures, such as LSTM-Autoencoders (LSTM-AE) or Transformer-Autoencoders, was considered but deemed unfeasible for this specific research phase. These models require high-cadence, uniformly sampled time-series data to effectively learn the latent temporal structure of an orbit. The current dataset, derived from operational ESA telemetry with a median cadence of 7200.00s and significant variable gaps, lacks the sequential density necessary to train a robust reconstruction-based temporal models without introducing significant interpolation artifacts. This architectural "blindness" to sequential continuity in the dense models is what necessitated the physics validation gate to act as the deterministic temporal link, filling the gap left by the lack of inherent recurrent units.

The failure of the Autoencoder family is rooted in the "contamination" of the latent space during training. Unlike supervised models that benefit from class balance, unsupervised Autoencoders require an overwhelming majority of "normal" data to establish a sharp reconstruction boundary. In this study, the 2:1 normal:anomaly ratio resulted in significant latent space contamination. Because anomalies represented over 30.00% of the training data, the bottleneck layer was forced to learn the compressed features of both normal and anomalous states. Consequently, the model became "too proficient" at reconstructing attacks, leading to the observed high FPRs (up to 90.26%) and an inability to distinguish malicious "jumps" from valid orbital drift.

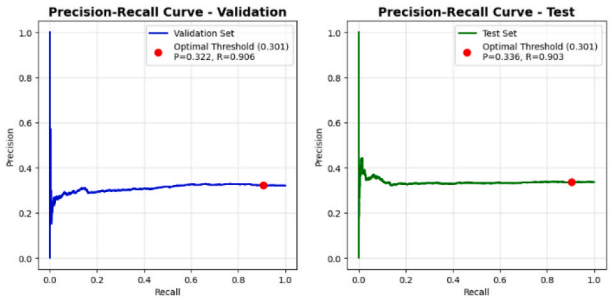

Fig. 12. PR curves for the autoencoder high recall focused model with a precision CI [33.24%, 34.00%] and recall CI [89.22%, 91.48%].

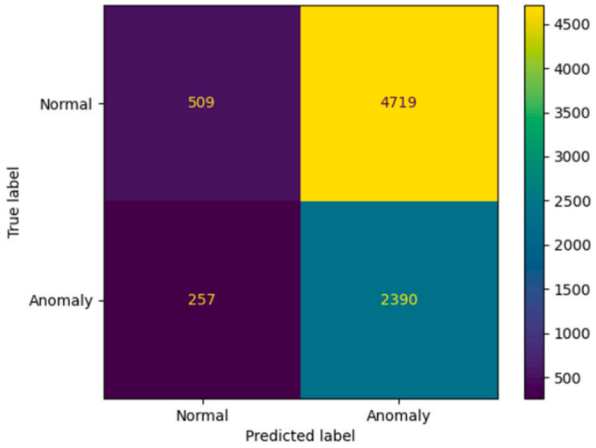

Fig. 13. Confusion matrix for the Test Set of the Autoencoder High Recall Focused Model.

While sophisticated architectures like LSTM-AE or Transformer-Autoencoders could theoretically capture temporal dependencies, they were not pursued due to an inherent objective mismatch. An Autoencoder focuses on $X \rightarrow Y$ (reconstruction), whereas orbital manipulation detection is fundamentally a $X_t \rightarrow Y$ (sequence-to-label) problem. In a high-stakes Ground Control Systems (GCS) environment, a predictive LSTM that focuses on the probability of an attack is architecturally superior to a reconstructive model that might "successfully" reconstruct a malicious state and thus fail to trigger an alert.

B. Overall Model Performance

When comparing the best models from each model family, it can be seen that LSTMs represent the optimal architecture because orbital manipulation attacks fundamentally exhibit temporal patterns that only sequential models can effectively capture. The high recall achieved by the balanced precision-recall model of LSTM suggests that it successfully learnt the underlying orbital manipulation signatures embedded in time-series orbital data.

Similarly, Random Forest and XGBoost models also serve as viable alternatives when interpretability requirements prevent deep learning, though it sacrifices some temporal modeling capabilities.

However, when considering the Autoencoder models, their approach fundamentally misaligns with the problem structure since in this case, Autoencoders apply unsupervised methods to an inherently supervised task.

The results of these models strongly indicate that temporal pattern recognition is crucial for this cybersecurity application, making sequential models like LSTMs the clear architectural choice for operational deployment. Crucially, the stark contrast between the high FPRs of the other models (up to 97.40%) when compared to the LSTM models (down to 4.84%) highlights the primary challenge: high FPRs translate directly into massive operational overhead and alert fatigue in a satellite control center, making the model untrustworthy. The physics validation gate, discussed in detail in Sub-Section C, directly addresses this operational necessity by converting statistical alerts into high-confidence

physical proofs, thereby significantly improving the trustworthiness of the entire system.

While the current models are optimized for "loud" kinematic footprints defined by significant deviations in orbital energy, the proposed framework establishes the foundation for detecting stealthier, unauthorized drifts. In a "stealth" attack scenario, an adversary might attempt to stay within the bounds of physical plausibility to evade the physics gate. However, by integrating sequential models like LSTMs that capture subtle temporal dependencies, the system is designed to identify the cumulative statistical signature of unauthorized maneuvers. As higher-cadence telemetry becomes available, the adaptive energy tolerance (51,000.00 J/kg) can be further tightened to transition the physics gate from gross anomaly detection to high-precision integrity monitoring, effectively closing the window for stealthy exploitation. The reliance on a kinematic proxy ensures that the system monitors the outcome of an attack rather than the vector, meaning that even a stealthy attack must eventually result in an unauthorized state change that the temporal ML layer is tuned to observe.

C. Physics Validation Results

This physics validation gate developed under this module is an excellent component to help identify the signature of a cyberattack designed for satellite orbital manipulation because it enforces fundamental physical constraints that injected data by an attacker cannot easily satisfy. As such, it identifies the attack by proving that the reported orbital change is kinematically inconsistent.

Satellite orbital manipulation attacks often require the satellite to instantly jump to a new, desired orbit or trajectory. This component uses the Tsiolkovsky Rocket Equation Principle to compare the required change in velocity in the telemetry against the known hardware limit of the satellite. An attacker attempting a large, instantaneous change (e.g., to hide the satellite or place it near a target) might inject data that implies a larger change in velocity and the physics validation gate immediately flags this as a Tsiolkovsky Rocket Equation Violation, proving the state change was not caused by the thrusters of the satellite, but by compromised data.

The most definitive proof of a satellite orbital manipulation attack induced by a cyberattack lies in the violation of energy conservation. The component uses the Vis-Viva Equation to check the change in specific orbital energy against the defined strict, adaptive tolerance. This energy is a function of both position and velocity. An attacker trying to spoof the location of a satellite must inject both a new, false position and a new, false velocity that matches the desired fake orbit. It is extremely difficult to calculate the perfect, precise, physically consistent position and velocity pair that satisfies the Vis-Viva Equation and links back to the previous state with a realistic maneuver. Since attackers typically inject inconsistent or mathematically mismatched position and velocity values, the multi-gate physics architecture captures these discrepancies.

The multi-gate physics architecture of this research acts as a non-negotiable, post processing filter that uses the laws of orbital mechanics to discard statistically flagged alerts that are nonetheless physically plausible. A false positive occurs when the ML model predicts an anomaly for an event that is actually a normal, non-anomalous event. Since the ML model does not understand physics, it cannot tell the difference between a high statistical deviation and a legitimate event. The physics gate subjects every alert to the laws of motion using the specified adaptive tolerance of 51,000.00 J/kg and calculates the change in energy to determine and discard false positives. By discarding alerts that are statistically abnormal but physically normal, the component directly reduces the false positive count.

The practical significance of the physics validation gate is best understood through the lens of mission operations. In a high-stakes SOC environment, a system with a FPR exceeding 90.00% as observed in the standalone Random Forest and XGBoost baseline models, is operationally unusable. Such high error rates lead to "alert fatigue", where human operators eventually ignore or disable security systems to maintain mission continuity. By integrating the physics gate, these probabilistic ML alerts are converted into high-confidence physical proof. This hybrid approach ensures that when an alert is escalated to a human operator, it has already passed a deterministic "Physical Law" test, thereby resolving the operational overhead challenges and establishing the system trustworthiness required for real-world deployment.

By removing false positives, the physics component inherently increases the final precision of the alerting system which can be calculated as follows:

$$Improved\ Precision = \frac{TP_{ML}}{\left(TP_{ML} + \left(FP_{ML} - FP_{Physics}\right)\right)} \tag{10}$$

- TP_{ML} refers to the true positives generated by the ML component.
- FP_{ML} refers to the false positives generated by the ML component.
- $FP_{Physics}$ refers to the false positives identified by the multi-gate physics architecture.

1) Random Forest Models

Table 16 showcases the test set performance of the physics validation gate for the Random Forest models. As can be seen, the physics gate of the baseline Random Forest model managed to further identify 519 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 33.92% by 3.46%, resulting in a final precision of 37.38%.

Furthermore, the physics gate of the high recall focused model managed to further identify 738 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 33.65% by 3.58%, resulting in a final precision of 37.24%.

Similarly, the physics gate of the balanced precision-recall model managed to further identify 412 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 33.37% by 3.29%, resulting in a final precision of 36.65%.

When considering the physics validation results of the Random Forest models, it can be seen that the most improvement in precision (3.58%) was observed for the best performing model of the Random Forest family – the high recall focused model. However, the best final precision of the three models was observed for the baseline Random Forest model (37.38%).

2) XGBoost Models

Table 17 showcases the test set performance of the physics validation gate for the XGBoost models. As can be seen, the physics gate of the baseline XGBoost model managed to further identify 230 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 32.31% by 3.14%, resulting in a final precision of 35.45%.

Similarly, the physics gate of the high recall focused model managed to further identify 728 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial

Table 16
Physics validation results for the random forest models.

Metric	Baseline	High Recall	Balanced Precision-Recall
Current Precision	33.92%	33.65%	33.37%
TP_{ML}	1901	2583	1534
FP_{ML}	3704	5092	3063
$FP_{Physics}$	519	738	412
Law of Energy Conservation Violations	5086	6937	4185
Improved Precision	37.38%	37.24%	36.65%
Improvement	3.46%	3.58%	3.29%

Table 17
Physics validation results for the XGBoost models.

Metric	Baseline	High Recall	Balanced Precision-Recall
Current Precision	32.31%	33.60%	33.10%
TP_{ML}	839	2536	1549
FP_{ML}	1758	5012	3131
$FP_{Physics}$	230	728	426
Law of Energy Conservation Violations	2367	6820	4254
Improved Precision	35.45%	37.18%	36.41%
Improvement	3.14%	3.59%	3.31%

precision of 33.60% by 3.59%, resulting in a final precision of 37.18%.

Furthermore, the physics gate of the balanced precision-recall model managed to further identify 426 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 33.10% by 3.31%, resulting in a final precision of 36.41%.

When considering the physics validation results of the XGBoost models, it can be seen that the most improvement in precision (3.59%) was observed for the best performing model of the XGBoost family – the high recall focused model. Similarly, the best final precision of the three models was also observed for the high recall focused XGBoost model (37.18%).

3) LSTM Models

Table 18 showcases the test set performance of the physics validation gate for the LSTM models. As can be seen, the physics gate of the baseline LSTM model managed to further identify 60 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 83.69% by 1.71%, resulting in a final precision of 85.40%.

Likewise, the physics gate of the high recall focused model managed to further identify 28 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 89.40% by 0.90%, resulting in a final precision of 90.30%.

Furthermore, the physics gate of the balanced precision-recall model managed to further identify 28 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 90.88% by 0.93%, resulting in a final precision of 91.81%.

When considering the physics validation results of the LSTM models, it can be seen that the most improvement in precision (1.71%) was observed for the baseline model of the LSTM family. However, the best final precision of the three models was observed for the best performing model of the LSTM family – the balanced precision-recall model (91.81%).

Furthermore, to address potential concerns regarding the high precision (90.88%) and the low FPR (4.84%) observed in the balanced precision-recall LSTM model, it is critical to distinguish between the performance of the ML component and the integrated hybrid system. As demonstrated by the baseline Random Forest and XGBoost results (Tables 12 and 13), traditional ML models alone produced high FPRs (up to 97.40%), reflecting the inherent noise and complexity of satellite telemetry. As such, the optimized final metrics are not a product of the ML layer alone but are achieved through the cascading architecture of the system. By tuning the ML layer for maximum recall and utilizing the physics validation gate as a deterministic filter for orbital mechanics violations (e.g., Vis-Viva and Tsiolkovsky constraints), the system achieves a validated reduction of false alarms that is grounded in physical reality rather than statistical probability.

Table 18
Physics validation results for the LSTM models.

Metric	Baseline	High Recall	Balanced Precision-Recall
Current Precision	83.69%	89.40%	90.88%
TP_{ML}	2510	2523	2521
FP_{ML}	489	299	253
$FP_{Physics}$	60	28	28
Law of Energy Conservation Violations	2939	2794	2746
Improved Precision	85.40%	90.30%	91.81%
Improvement	1.71%	0.90%	0.93%

4) Autoencoder Models

Table 19 showcases the test set performance of the physics validation gate for the Autoencoder models. As can be seen, the physics gate of the baseline Autoencoder model managed to further identify 383 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 33.89% by 3.78%, resulting in a final precision of 37.67%.

Furthermore, the physics gate managed to further identify 383 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 32.97% by 4.92%, resulting in a final precision of 37.89%.

Similarly, the physics gate managed to further identify 383 false positive cases out of the total number of anomalies detected by the ML model. This helped to boost the initial precision of 33.62% by 1.91%, resulting in a final precision of 35.53%.

When considering the physics validation results of the Autoencoder models, it can be seen that the most improvement in precision (4.92%) was observed for the VAE model of the Autoencoder family. This was also the highest improvement in precision across any model that was observed during results analysis. Similarly, the best final precision of the three models was also observed for the VAE model (37.89%).

Another notable observation in the physics validation results of the Autoencoder models is the recurrence of identical false positive reduction values across the three model variants. This consistency is not a clerical error but a fundamental property of the system's architecture. The physics validation gate operates as a deterministic "sieve" with fixed physical tolerances (e.g., the 51,000.00 J/kg energy limit). While different ML models (the "nets") may flag varying sets of data points as statistical anomalies, the subset of those alerts that represent physically plausible noise and are thus filtered out, is determined solely by the inherent physical properties of the telemetry. Consequently, for model families with similar detection boundaries, the gate consistently identifies and removes the same underlying physical artifacts, demonstrating its reliability as an independent integrity check.

D. Overall Physics Validation Results

When considering the results of the multi-gate physics architecture, it can be seen that this was a very efficient approach of improving the precision and reducing the number of false positives of the system without compromising the primary metric of the research – the recall. As can be seen, all best performing model variants surpassed the defined threshold for precision with the precision of the LSTM model far

Table 19
Physics validation results for the autoencoder models.

Metric	Baseline	VAE	High Recall
Current Precision	33.89%	32.97%	33.62%
TP_{ML}	1293	973	2390
FP_{ML}	2522	1978	4719
$FP_{Physics}$	383	383	383
Law of Energy Conservation Violations	3432	2568	6726
Improved Precision	37.67%	37.89%	35.53%
Improvement	3.78%	4.92%	1.91%

surpassing the other models, emerging as the top contender for an operationally ready model. Table 20 summarizes the comparison of the precision improvements of the best performing model variants before and after applying the physics validation gate.

A comparative breakdown of the physics validation results reveals an inverse relationship between initial ML performance and the gate's relative impact. For model families with higher initial FPRs, such as Random Forest and Autoencoders, the physics gate acted as a robust filter, identifying hundreds of kinematically inconsistent anomalies and resulting in precision gains as high as 3.58% and 4.92% respectively. Conversely, for the highly optimized LSTM models, which already possessed high architectural precision, the gate provided a smaller but critical precision boosts (from 0.90% to 1.71%). This demonstrates that the physics gate is most impactful as a "safety net" for traditional ML models, while serving as a high-fidelity verification layer for more advanced temporal architectures.

It is critical to differentiate between physical impossibility and operational unauthorized action. The Vis-Viva gate acts as a deterministic filter for gross physical violations (such as sensor spoofing where energy is not conserved). Conversely, the ML layer is tasked with identifying patterns that, while physically plausible, deviate from authorized mission norms. This dual-layered approach ensures that the system is not merely a physics-checker, but a comprehensive cybersecurity defense capable of identifying both data-level corruption and high-level behavioral manipulation.

E. Computational Complexity and Operational Feasibility

To assess the practical readiness of the proposed system for a real-world SOC, the computational costs were evaluated across training, inference, and physics-informed filtering stages.

1) Model Training Duration

Table 21 showcases the time taken to train each model variant. The tree-based model family XGBoost demonstrated the highest efficiency, training all its variants in under 4.00s. While the LSTM model variants required significantly more time (up to 632.29s for the baseline variant), they remain viable as training is typically an offline process performed during the mission preparation or update phases.

2) Inference and Physics Overhead

During the testing phase, individual inference times for all models including the deep learning-based LSTM were measured in the millisecond range per telemetry window. The physics validation gate introduced negligible computational overhead (<1.00 ms), as it involves deterministic algebraic calculations (Vis-Viva Equation) rather than iterative optimization. Consequently, the entire hybrid pipeline can process incoming telemetry in near real-time, comfortably fitting within the standard polling intervals of modern satellite GCS.

Table 20
Comparison of precision improvements of the best performing model variants before and after applying the physics validation gate.

Metric	Random Forest	XGBoost	LSTM	Autoencoder
Precision Before Physics Gate	33.65%	33.60%	90.88%	33.62%
TP_{ML}	2583	2536	2521	2390
FP_{ML}	5092	5012	253	4719
$FP_{Physics}$	738	728	28	383
Law of Energy Conservation Violations	6937	6820	2746	6726
Precision After Physics Gate	37.24%	37.18%	91.81%	35.53%
Improvement	3.58%	3.59%	0.93%	1.91%

Table 21
Time taken to train each model.

Model Family	Variant	Training Time
Random Forest	Baseline	27.59s
	High Recall	26.34s
	Balanced Precision-Recall	31.03s
XGBoost	Baseline	2.55s
	High Recall	3.66s
	Balanced Precision-Recall	2.88s
LSTM	Baseline	632.29s
	High Recall	312.34s
	Balanced Precision-Recall	188.46s
Autoencoder	Baseline	49.46s
	VAE	416.84s
	High Recall	119.95s

From an operational perspective, the minimal computational footprint of the inference and physics validation stages (<1.00 ms overhead) suggests that the system could be deployed on standard GCS hardware without requiring specialized High-Performance Computing (HPC) clusters. Because the system processes telemetry in discrete windows (±7200.00s), the computational load is periodic rather than continuous, allowing it to run as a background service alongside existing flight dynamics and housekeeping monitors. This low-latency profile is essential for "Human-in-the-loop" operations, ensuring that orbital integrity alerts reach operators fast enough to allow for command-level intervention.

3) Computational Latency and Real-Time Feasibility

The physics validation layer was benchmarked to evaluate its suitability for high-cadence GCS. The observed latency of <1.00 ms refers to the per-record processing time for the Vis-Viva energy calculation. All benchmarks were conducted on a workstation equipped with an AMD Ryzen 5 7530U with Radeon Graphics (2.00 GHz) processor and 32.00 GB RAM. The system was developed in Python 3.13 (leveraging the latest interpreter performance optimizations) using NumPy for vectorized mathematical operations and Pandas for high-speed data orchestration. The <1.00 ms overhead includes the time required to retrieve the state vector, calculate the specific orbital energy, and compare it against the calibrated tolerance (51,000.00 J/kg).

6. Limitations

This research represents a crucial first step in a domain lacking exhaustive empirical data. The use of public orbital anomaly data as a kinematic proxy was a methodological necessity; however, it inherently biases the models toward "loud" kinematic footprints which can be defined as major, sudden deviations in orbital state. This leaves a vulnerability to stealthy manipulation attacks, where an adversary might slowly drift a satellite over several days. In such cases, the manipulation might remain within "physically plausible" bounds, potentially bypassing the deterministic physics gate. The technical roadmap to address these stealthier threats involves two primary advancements.

- Temporal Accumulation Analysis: Future iterations will move beyond instantaneous window checks (±7200.00s) to multi-day trend analysis. By training LSTMs or Transformers on longer sequences, the system can detect the cumulative statistical "drift" of an unauthorized maneuver that is individually small but collectively significant.
- Dynamic Tolerance Tightening: As telemetry cadence moves from hours to minutes, the adaptive energy tolerance (currently 51,000.00 J/kg) can be tightened using high-fidelity perturbation modeling (e.g., J_2 through J_4 gravity models). This reduces the

"search space" available for an attacker to hide a malicious maneuver within natural orbital noise.

By evolving the system to monitor the rate of change of energy rather than just absolute discrepancies, the framework can transition from detecting gross physical impossibilities to identifying subtle, unauthorized behavioral shifts.

The resulting models of most model families, though achieving high recall for statistical outliers, were plagued with high false positives that suggest that they were detecting general orbital deviations rather than specific cyberattack induced deviations. The physics validation gate which was deployed to address these false positives deemed successful up to a certain extent, improving the precision of these models without compromising the recall. Furthermore, while the integrated model provides a strong proof of concept on this dataset, its true generalizability to other satellite missions with different hardware and operational profiles remains unproven.

When considering the results of the multi-gate physics architecture, it was observed that there were zero Tsiolkovsky Rocket Equation violations in the test sets. The limitation imposed by this observation is that the resilience of the current integrated system against a key class of cyberattack remains unvalidated. The existing dataset only contains anomalies characterized by energy inconsistency and not those requiring a kinematically inconsistent change in velocity. In the long term, transparency through public disclosure of anonymized, sanitized telecommand attack data is the most effective and representative way to validate the full spectrum of space cybersecurity defenses, including the critical Tsiolkovsky Rocket Equation gate. Having access to such data would ensure that the Rocket Equation gate is functionally validated and that the system is fully robust against diverse orbital manipulation attempts.

Proxy mode, as defined in this research, referred to the operation configuration where the system uses simplified, scalar features instead of the full 3D vector telemetry for physics calculations. The Angular Momentum gate acts as a crucial check against orbital plane manipulation, but it remained unvalidated in this research because the system was restricted to proxy mode. This check relies on the principle that the direction of the specific angular momentum vector must be highly conserved unless a significant external force acts perpendicular to the orbital plane. The gate measures the angle change between the initial and final vectors, comparing it to a very tight tolerance. For instance, a major plane change requires a massive change in velocity, amounting to hundreds or thousands of m/s, often exceeding the defined threshold. An attacker who changes the reported inclination without injecting the massive corresponding change in velocity will cause an angular momentum violation. This proves that the orientation data of the satellite was spoofed, identifying the attack. However, since the proxy data only contained the magnitudes of position and velocity, and not the full three-dimensional vectors required to compute the vector cross-product, the gate could not be executed, leaving the resilience of the system to inclination spoofing attacks unverified.

In essence, this research successfully highlights the potential for AI in space cybersecurity while also clearly defining the significant challenges that must be overcome before these models can be safely and effectively deployed on operational satellites. Primarily, it highlights the need for publicly available telecommand cyberattack data to validate the Tsiolkovsky Rocket Equation gate, and the requirement to move beyond proxy mode to verify the crucial Angular Momentum gate against orbital plane manipulation.

7. Future research

This section discusses three major concerns stemming from current limitations to focus on optimizing the ML detection layer to complement the high-confidence filtering provided by the physics validation gate.

A. Advanced Decision Making and Thresholding
- Implementation of Multi-Criteria Decision Analysis (MCDA) for Alert Triage: This approach helps to combine multiple validation metrics into a structured decision framework, thereby replacing simple binary classification and allowing for sophisticated alert triage.
- Implementation of Dynamic Thresholding with Confidence Scoring: This approach helps to replace the current fixed statistical threshold with adaptive, confidence-based decision boundaries and is crucial for improving the detection of anomalies that are just below the defined thresholds.

B. Improving Temporal Anomaly Detection
- Implementation of Temporal Pattern Enhancement with Attention Mechanisms: This approach helps to modify the existing LSTM architectures to better process and focus on anomaly-relevant patterns within the low-cadence telemetry gaps. This addresses the inherent difficulty of detecting subtle, complex attacks when data points are widely spaced in time.
- Benchmarking with Advanced Sequential Models: Future work could involve benchmarking against more recent and sophisticated temporal models, such as Transformer variations or other hybrid sequential architectures. This will provide richer insights into maximizing performance for identifying complex, time-linked anomaly patterns.

C. Empirical Verification of Design-Stage Physics Gates and Enhancing Reliability and Generalizability
- Validation of Tsiolkovsky Gate Using Synthetic Data: While the Tsiolkovsky Rocket Equation gate is architecturally integrated into the current system design, this high-priority step involves generating synthetic attack data that requires a kinematically inconsistent change in velocity. This will formally validate the Tsiolkovsky Rocket Equation gate, ensuring the system is robust against this primary class of orbital manipulation attacks to empirically verify its performance against velocity-jump attacks. This gate requires time-correlated fuel mass flow rates (dm/dt) and onboard propellant tank pressure/temperature. While the Vis-Viva gate detects if an energy change occurred, the Tsiolkovsky gate validates how it occurred. It is specifically designed to detect "Velocity-Jump" anomalies, where a reported Δv is kinematically inconsistent given the satellite's remaining mass and thruster specific impulse (I_{sp}), even if the resulting state is energy-consistent.
- Validation of Angular Momentum Gate and Full-Vector Telemetry: To verify the resilience of the system against orbital plane manipulation (inclination spoofing), future work must move beyond the current proxy mode. This requires acquiring or simulating full 3D cartesian position and velocity vectors to calculate the specific angular momentum vector, thereby enabling the crucial Angular Momentum gate to check for non-Keplerian plane changes (violations of the Law of Conservation of Angular Momentum). This gate requires full 3D cartesian state vectors to calculate the specific angular momentum vector. Its primary role is to detect "energy-neutral plane changes". For instance, an adversary could spoof an inclination or node change (orbital plane shift) that keeps the scalar orbital energy constant. The Vis-Viva gate is blind to these orientation-only attacks, whereas the Angular Momentum gate would immediately flag the violation of the conservation of the orbital plane.
- Implementation of Ensemble Methods with Diversity-Based Selection: This approach helps to create intelligent ensembles that combine models trained on different features to leverage complementary strengths and would specifically target the reduction of the remaining false positives by requiring multiple, diverse models to agree on an alert.
- Establishing Pathways for Representative Attack Data: It is imperative that a commitment to exploring collaboration be made

with space agencies and industry partners to develop an ethical framework for the disclosure of anonymized, sanitized telemetry data related to confirmed cyber incidents. Access to such data is the most effective way to ensure the full spectrum of space cybersecurity defenses is validated against realistic threat scenarios.

Table 22 highlights the technical requirements to transition the remaining physics validation gates from an architectural design state to empirical validation. Future validation of these gates requires moving beyond current "proxy mode" telemetry to incorporate high-fidelity 3D vectors and propulsion-specific logs.

D. Operational Integration Deployment
- Investigation of Real-Time Computational Constraints and Integration with GCS: A critical next step is to investigate the real-time processing latency and computational requirements of this proposed physics-informed ML approach. This includes designing Application Programming Interfaces (APIs) and data interfaces to facilitate seamless integration with existing satellite GCS to ensure operational readiness and efficient alert triage by human operators.

E. Mitigating Stealthy and Physically Plausible Threats
- Multimodal Propellant-State Correlation: By cross-referencing detected orbital state changes with onboard propulsion telemetry (e.g., fuel tank pressure and temperature), the system can identify "ghost maneuvers". A state change that is physically possible according to Vis-Viva, but lacks a corresponding propellant-depletion signature, would be flagged as a high-confidence unauthorized injection.
- Differential Energy Gradient Analysis: Transitioning from static thresholds to monitoring the rate of change in specific energy over high-cadence intervals. By reducing the sampling window, the uncertainty contributed by natural perturbations (e.g., J_2, atmospheric drag) is minimized, allowing the system to detect subtle, cumulative energy "drifts" that characterize low-thrust stealth attacks.

8. Conclusion

Since concerns on cyberwarfare and space warfare are increasingly becoming important topics of discussion in the world today, especially in the field of satellites, the development of reliable, physics-based cybersecurity models for satellite systems is now a national security imperative. As such, this research proposed a novel physics-informed ML approach to detect suspicious orbital maneuvers. This study made use of a custom telecommand-related dataset derived from the ESA-ADB dataset by filtering telemetry data to focus on anomalies occurring within ± 48.00 h of command executions. This temporal filtering approach resulted in a naturally balanced dataset. The findings of this research successfully established that temporal pattern recognition is paramount for detecting satellite orbital manipulation attacks, with LSTM networks emerging as the most promising model architecture. By leveraging their ability to learn sequential dependencies, the LSTM models achieved a good recall rate of 95.64% with a similarly high precision of 90.88%, demonstrating their superior capability for identifying threats that unfold over time. While Random Forest and XGBoost models showed strong secondary performance, Autoencoders proved to be fundamentally unsuitable for this application. This underperformance is attributed to their architectural nature as point-in-time reconstructors; without recurrent units or attention mechanisms, they treat telemetry snapshots as isolated events, failing to model the kinematic continuity of an orbit. Furthermore, the balanced nature of the filtered dataset (2:1 normal:anomaly ratio) prevented the Autoencoders from establishing a sufficiently sharp "normal" reconstruction boundary, leading to the observed high FPRs. The novel component introduced in this research, the multi-gate physics architecture, contributed towards reducing the number of false positives and improving the precision of the model without compromising the recall. This component is also vital because it acts as the final, non-negotiable proof, transforming raw statistical alerts into high-confidence cybersecurity indicators by confirming that detected anomalies are kinematically inconsistent events that violate the fundamental laws of orbital mechanics. Specifically, the results validate the effectiveness of the energy-based Vis-Viva gate in reducing false positives, while providing a designed framework for Tsiolkovsky and Angular Momentum validation as the availability of high-fidelity 3D vector and thrust-specific cyberattack data increases. This crucial filtering step dramatically improves the trustworthiness of the entire system, making it a robust proof of concept nearing operational readiness.

Table 22
Technical requirements to transition the remaining physics modules from architectural design to empirical validation.

Physics Gate	Primary Data Requirements	Targeted Anomaly (Missed by Vis-Viva)
Tsiolkovsky	Effective exhaust velocity (v_e), initial mass (m_i), and final mass (m_f)	Velocity-Jump Attacks: Detects propulsive maneuvers exceeding physical hardware limits, even if the reported position/velocity pair is energy-consistent
Angular Momentum	Full 3D cartesian position (r) and velocity (v) vectors	Orbital Plane Manipulation: Detects unauthorized changes in inclination or node (plane-shifting) where the scalar energy remains constant, but the vector orientation is spoofed

CRediT authorship contribution statement

K.K.H. Karunathilake: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kavinga Yapa Abeywardena:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Conceptualization. **Sara Vecchini:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization.

Funding

The authors declare that no external funding, grants, or financial support was received for the research and authorship of this article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Mr. Prasanna Shankar who is a part of the Ansys/Synopsys Academic Program Team for his support in obtaining the required licensing for Ansys Systems Tools Kit (STK), which was extremely beneficial for the research. The authors would also like to express their gratitude to all reviewers and editors for their valuable feedback in improving this publication.

Data availability

Data will be made available on request.

References

[1] Maral G. Satellite communication systems. John Wiley & Sons Ltd.; 2020.
[2] Laursen L. Satellite Signal Jamming Reaches New lows - Starlink and Other LEO Constellations Face a New Set of Security Risks. IEEE Spectrum; 2023 [Online]. Available: https://spectrum.ieee.org/satellite-jamming [Accessed 15 February 2025].
[3] O'Neill PH. Russia Hacked an American Satellite Company One Hour before the Ukraine Invasion. MIT Technology Review; 2022 [Online]. Available: https://www.technologyreview.com/2022/05/10/1051973/russia-hack-viasat-satellite-ukraine-invasion/ [Accessed 15 February 2025].
[4] Thummala R, Falco G. Hacktivism goes orbital: investigating NB65's breach of ROSCOSMOS. AIAA SciTech Forum; 2024.
[5] Pavur J, Martinovic I. Building a launchpad for satellite cyber-security research: lessons from 60 years of spaceflight. Journal of Cybersecurity 2022;8.
[6] Hamill-Stewart J, Rashid A. Threats against Satellite ground infrastructure: a retrospective analysis of sophisticated attacks. In: Workshop on security of space and satellite systems (SpaceSec) 2024, San Diego; 2024.
[7] Willbold J, Schloegel M, Vögele M, Gerhardt M, Holz T, Abbasi A. Space odyssey: an experimental software security analysis of satellites. In: IEEE symposium on security and privacy (SP); 2023.
[8] Jones RC. Survey of space Professionals' Perception of Satellite Cybersecurity from 2012 to 2022: decision-makers' thoughts on satellite cybersecurity evolving. New Space 2024;12(4):259–72.
[9] Gough E. A Chinese Space Tug Just Grappled a Dead Satellite. Universe Today; 2022 [Online]. Available: https://www.universetoday.com/articles/a-chinese-space-tug-just-grappled-a-dead-satellite [Accessed 19 August 2025].
[10] Falco G, Thummala R, Kubadia A. WannaFly: an approach to satellite ransomware. In: 2023 IEEE 9th international conference on space Mission challenges for information technology (SMC-IT); 2023.
[11] Benecki P, Piechaczek S, Kostrzewa D, Nalepa J. Detecting anomalies in spacecraft telemetry using evolutionary thresholding and LSTMs. In: Gecco '21: genetic and Evolutionary Computation Conference; 2021.
[12] Jin X, Wang HQ, Jin ZH. Anomaly detection of satellite telemetry data based on extended dominant sets clustering. J Phys Conf 2023;2489.
[13] Baireddy S, Desai SR, Methieson JL, Foster RH, Chan MW, Comer ML, Delp EJ. Spacecraft time-series anomaly detection using transfer learning. In: 2021 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW); 2021.
[14] Wang Y, Gong J, Zhang J, Han X. A deep learning anomaly detection framework for satellite telemetry with fake anomalies. International Journal of Aerospace Engineering 2022;2022.
[15] He J, Cheng Z, Guo B. Anomaly detection in satellite telemetry data using a sparse feature-based method. Sensors 2022;22.
[16] Kricheff S, Maxwell E, Plaks C, Simon M. An explainable machine learning approach for anomaly detection in satellite telemetry data. In: 2024 IEEE aerospace conference; 2024.
[17] Habib BE, Boualem N. Detection and prediction of satellite telemetry anomalies. In: 2024 4th international conference on Embedded & Distributed Systems (EDiS); 2024.
[18] Zeng Z, Jin G, Xu C, Chen S, Zhang L. Spacecraft telemetry Anomaly Detection based on parametric causality and double-criteria drift streaming peaks over threshold. Appl Sci 2022;12.
[19] Ali MAH. A real-time Anomaly detection in satellite telemetry data using artificial intelligence techniques depending on time-series analysis. Journal of the ACS Advances in Computer Science 2023;14.
[20] Bieber M, Verhagen WJC, Cosson F, Santos BF. Generic diagnostic framework for Anomaly detection—application in satellite and spacecraft systems. Aerospace 2023;10(8):673.
[21] Hu F, Wang Q, Shao H, Gao S, Yu H. Anomaly detection of UAV state data based on single-class triangular global alignment kernel extreme learning machine. Comput Model Eng Sci 2023;136(3):2405–25.
[22] Feng C, Fan J, Liu Z, Jin G, Chen S. Unmanned aerial vehicle anomaly detection based on causality-enhanced graph neural networks. Drones 2025;9(6):408.
[23] Huakun C, Yongxi L, Jingping S, Weiguo Z. UAV anomaly detection method based on convolutional autoencoder and support vector data description with 0/1 soft-margin loss. Drones 2024;8(10):534.
[24] Bell V, Arce IM, Mase JM, Rengasamy D, Rothwell B, Figuerdo GP. Anomaly detection for unmanned aerial vehicle sensor data using a stacked recurrent autoencoder method with dynamic thresholding. SAE International Journal of Aerospace 2022;15(2):219–29.
[25] Chen S, Jin G, Long X. On-orbit satellite hierarchical anomaly detection using causal structure learning. Adv Space Res 2025;75(1):718–36.
[26] Cui L, Zhang Q, Shi Y, Yang L, Wang Y, Wang J, Bai C. A method for satellite time series anomaly detection based on fast-DTW and improved-KNN. Chin J Aeronaut 2023;36(2):149–59.
[27] Song Z, Mu Z, Wu S, Jin S, Yi J. Ault diagnosis of spacecraft electrical power system based on improved Newman community divisions method. Chin J Aeronaut 2025:103752.
[28] Yang K, Wang Y, Han X, Cheng Y, Guo L, Gong J. Unsupervised anomaly detection for time series data of spacecraft using multi-task learning. Appl Sci 2022;12(13).
[29] Cuéllar S, Santos M, Alonso F, Fabregas E, Faria G. Explainable anomaly detection in spacecraft telemetry. Eng Appl Artif Intell 2024;133:108083.
[30] Sutton GP, Biblarz O. Rocket propulsion elements. Wiley; 2001.
[31] Falco G. When Satellites attack: satellite-to-Satellite cyber attack, defense and resilience. In: Ascend; 2020.
[32] Aderinto AB. "Advanced Cybersecurity Frameworks for Protecting Satellite Networks, Deep-Space Communications, and Space Assets". International Journal of Research Publication and Reviews 2025;6(2):4729–45.
[33] Kotowski K, Haskamp C, Andrzejewski J, Ruszczak B, Nalepa J, Lakey D, Collins P, Kolmas A, Bartesaghi M, Martinez-Heras J, De Canio G. European space agency benchmark for Anomaly detection in satellite telemetry. ArXiv 2024;2406:17826.
[34] Kotowski K, Haskamp C, Andrzejewski J, Ruszczak B, Nalepa J. Annotating large satellite telemetry dataset for ESA international AI anomaly detection benchmark. In: Big data from space (BiDS'23); 2023.
[35] Hossain MA. Orbital decay of low Earth orbit satellites: a numerical investigation. International Journal of Scientific Engineering and Science 2023;7(8):24–34.
[36] Nelkon M, Parker P. Advanced level physics. Heinemann; 1995.
[37] Hay A. The tsiolkovsky rocket equation: a parallel derivation. Principium; May 2022.
[38] Curtis HD. Orbital mechanics for engineering students. Elsevier Butterworth-Heinemann; 2005.
[39] Vallado DA. Fundamentals of astrodynamics and applications. Springer; 2013.
[40] Software for digital mission engineering and systems analysis," Ansys STK, [Online]. Available: https://www.ansys.com/products/missions/ansys-stk. [Accessed 19 October 2025].