AI SSA Challenge Problem: Satellite Pattern-of-Life Characterization Dataset and Benchmark Suite

Peng Mun Siew[*], **Haley E. Solera**[†], **Thomas G. Roberts**[‡], **Daniel Jang**[§]
Massachusetts Institute of Technology
Victor Rodriguez-Fernandez[¶]
Universidad Politécnica de Madrid

Jonathan P. How[‖], **Richard Linares**[**]
Massachusetts Institute of Technology

ABSTRACT

With the increasing number of resident space objects (RSOs), it has become imperative to enhance tracking and orbit prediction capabilities to safeguard space assets from the threat of object-on-object collision. One effective approach to achieve this goal is by characterizing active satellites' patterns of life (PoL), or the sequences of behavioral modes—periods of consistent on-orbit behavior, such as those in which satellites adhere to various station-keeping protocols—that they pursue throughout their operational lifetimes. Understanding a satellite's PoL can contribute to better orbit prediction and improved tracking capabilities. In this work, a novel benchmarking tool for geosynchronous satellite pattern-of-life characterization was created to compare different AI approaches developed by independent research teams that push the boundary of this problem. To this end, the author team has developed a labeled Satellite Pattern-of-Life Identification Dataset (SPLID) containing astrometric data and satellites' historical behavioral modes with which AI-enabled algorithms can characterize satellite PoLs. The SPLID dataset consists of synthetic space object data, true space object data generated from Vector Covariance Messages (VCMs), and true space object data generated from high-accuracy ephemerides provided by satellite owner-operators.

In addition, this paper introduces a new AI SSA challenge competition that will be held in conjunction with the release of the SPLID dataset, where participants are tasked to build models that accurately label and time stamp the behavioral modes of GEO satellites over a six-month period. The satellite PoL characterization problem poses a unique set of challenges that combines complex multivariate time-series data analysis, behavioral pattern recognition, change point detection, and anomaly detection. The complexity of the problem, its real-world applicability and the potential impact of improving SSA makes this unique challenge an excellent platform for research, experimentation, and education. Furthermore, from the point of view of AI research, this challenge is also compelling due to the relative scarcity of public benchmarks involving time series data, especially those that require multivariate modeling and anomaly change point detection over long, noisy sequences. Pushing the limits on this proposed task could advance core time series capabilities and inspire innovative solutions transferable to other domains.

A machine learning algorithm and a heuristic-based approach are implemented to serve as baseline methods for the challenge problem. These baselines will aid in assessing the performance of more advanced AI solutions and help participants in getting started on the challenge problem. The baseline machine learning implementation will be provided along with a development kit to assist participants in familiarizing themselves with the data, methodologies, and evaluation criteria. The dataset and baseline solutions are available online at: `arclab.mit.edu/aichallenge`

[*]Research Scientist, Department of Aeronautics and Astronautics. E-mail: siewpm@mit.edu
[†]Ph.D. Candidate, Department of Aeronautics and Astronautics. E-mail: hsolera@mit.edu
[‡]Ph.D. Candidate, Department of Aeronautics and Astronautics. E-mail: thomasgr@mit.edu
[§]Ph.D. Candidate, Department of Aeronautics and Astronautics. E-mail: djang@mit.edu
[¶]Associate Professor, Department of Computer Systems Engineering. E-mail: victor.rfernandez@upm.es
[‖]Professor, Department of Aeronautics and Astronautics. E-mail: jhow@mit.edu
[**]Associate Professor, Department of Aeronautics and Astronautics. E-mail: linaresr@mit.edu

1. INTRODUCTION

Despite the attention garnered in recent years by applying innovative artificial intelligence (AI) techniques to challenging problems in aeronautics and astronautics, there is still a lack of adoption among the space situational awareness (SSA) community. Although historical data on Earth-orbiting satellites—troves of orbital state and light curve measurements derived from passive observations of space objects from the ground- and space-based sensors—may be particularly well-suited for AI-driven analysis, the disconnect between the AI and SSA research communities have prevented robust, interdisciplinary research progress. AI algorithms excel at recognizing and learning from patterns and can assist in identifying subtle patterns and correlations that might not be apparent to human analysts. Besides that, manual analysis of the SSA data is becoming increasingly challenging with the increasing number of space objects and the vast number of data that are generated daily. These repetitive tasks can be easily offloaded to AI algorithms, freeing up important human resources for other tasks that require human expertise and decision-making. Furthermore, advanced AI algorithms can help predict future behaviors and assist in proactive decision-making for satellite operations.

Much of the AI research community lacks domain-specific technical knowledge and suffers from limited SSA data availability. The challenge problem introduced in this paper aims to leverage expertise from the AI research community to use multi-faceted SSA data for characterizing a satellite pattern-of-life (PoL) in new, innovative ways. The problem poses a unique challenge that combines complex multivariate time-series data analysis, behavioral pattern recognition, and anomaly detection while having a real-world impact in contributing toward a safe and sustainable space environment. In addition, satellite behaviors can be highly variable from one satellite to another due to a lack of common operational guidelines, mission objectives, and propulsion systems. This data variability further increases the difficulty of the problem, and makes manual analysis ineffective. AI techniques are needed to automatically find complex patterns, integrate diverse data sources, and uncover non-intuitive relationships obscured in noisy time series.

Furthermore, the challenge problem also serves as an excellent platform for research, experimentation, and education, allowing the AI community to collaborate with the space community and showcase its breakthrough to a wider community.

This AI SSA challenge problem focuses on a specific orbital regime; the Geosynchronous Earth Orbit (GEO). GEO satellites have an orbital period that matches the Earth's rotation period, resulting in the satellite returning to the same position in the sky after a period of one sidereal day for an observer on Earth's surface. This unique property of GEO makes it valuable for various applications, ranging from communication, weather monitoring, guidance and navigation, to earth observation.

Over the course of GEO satellites' operational lifetimes, operators issue commands to place them in various behavioral modes, ranging from station-keeping, to longitudinal shifts, to end-of-life behaviors, and perhaps many others. Satellite PoLs are descriptions of on-orbit behavior made up of sequences of both natural and non-natural behavior modes [1, 2].

For the majority of GEO satellites, station-keeping is the most commonly observed class of behavioral mode. Satellites perform station-keeping maneuvers to counteract external perturbations at play in that regime, including third body effects of the sun and the moon, the variation in the Earth's gravitational field due to the non-spherical and in-homogeneous mass distribution of Earth, and solar radiation pressure. Station-keeping maneuvers can be further classified into north-south and east-west station-keeping, referring to corrections to displacements in geographic latitude and longitude, respectively. The third body effects of the Sun and Moon cause a drift in the inclination of RSOs in GEO, requiring high-energy north-south corrections. On the other hand, the variation in the Earth's gravitational field and solar radiation pressure causes a drift in the RSOs' eccentricity. Due to Earth's non-uniform gravitational field, GEO RSOs tend to drift toward the GEO gravity wells located at 75.1°E and 105.3°W.

One common way for a GEO RSO to shift from one behavioral mode to another is to perform a longitudinal-shift maneuver, which changes the satellite's mean longitude in a controlled manner [3]. This type of maneuver is typically driven by changes in the satellite's mission objectives. Another time when satellites typically change their behavioral mode is at the end of their lifespan, when they may change station-keeping protocols to conserve fuel or perform retirement maneuvers, where the GEO satellite is maneuvered to a higher altitude, outside of the geostationary belt.

Identifying these behavioral modes—including those not associated with well-understood operational patterns—would assist the SSA research community in better understanding the behavior of satellites. This enhanced understanding

of the satellite behavior, in turn, leads to more accurate predicted trajectories, improved tracking capabilities, and enhanced assessments of satellite conjunction. Furthermore, characterizing the PoLs for a diverse range of GEO satellites can help to contextualize historic on-orbit behaviors and behavior patterns, cultivate generalized maneuver prediction on a large scale, enable early identification of future satellite behaviors, enable inference of metadata from new and historic behaviors, as well as enable better mission planning [4].

Artificial intelligence (AI) approaches have been shown to be able to learn the relationship between time-series data and have previously been successfully applied to complex multi-dimensional problems such as spacecraft anomaly detection [5], orbit estimation [6], or weather prediction (including space weather [7]), among others. In terms of algorithms, while deep learning has become the standard approach for language and vision tasks, its adoption for time series analysis has been slower and less pronounced. This lag in acceptance is due, in part, to the fact that accuracy improvements for common tasks like time series classification have not been as substantial [8]. However, recent years have seen the transfer of cutting-edge deep learning techniques, such as self-supervised methods and transformer-based architectures, to time series data [9]. One of the main challenges in developing AI approaches for this application is the difficulties in acquiring realistic state vector data that are of sufficient quality and temporal resolution. Furthermore, there is no common dataset making it impossible to evaluate and compare the performance of different AI algorithms. Over the last decade, several space-related challenge competitions have been held with success by the Advanced Concepts Team of the European Space Agency [10–13]. Through these challenges, they were able to motivate the adoption and development of AI approaches to the space community.

The Satellite Pattern-of-Life Identification Dataset (SPLID), created by the author team to support this challenge problem, consists of synthetic astrometric data generated using a high-fidelity simulator to simulate a range of operation scenarios. In addition to this public set of synthetic astrometric data, the challenge problem will also utilize a private set of real astrometric data for algorithm evaluation. The real astrometric data includes high-accuracy ephemeris data provided by satellite owners and operators as well as Vector Covariance Message (VCM) data that was provided by the United States Joint Space Operations Center (JSpOC).

A development kit [1] coded in both Matlab and Python consisting of basic utility functions for data parsing, manipulation, and visualization has been made available alongside the SPLID. Baseline machine learning solutions will also be provided. The baseline solutions will be coded in Python using readily available packages and aim to lower the barrier of entry to AI techniques in the SSA research community.

The challenge is hosted on the EvalAI website, an open-source platform to run end-to-end AI challenges at scale [14]. EvalAI provides automation for key features like participant submissions, evaluation metrics, live leaderboards, and discussion forums. The theme of the AI SSA challenge competition is automating PoL discovery for satellites within GEO using AI under sparse and degraded SSA data. Submissions to the AI SSA challenge competition will be evaluated based on their precision and recall scores.

The main contribution of this work is to introduce the SPLID dataset and the satellite PoL characterization challenge problem. Furthermore, two baseline solutions using an analytical algorithmic PoL node detection algorithm and machine learning are discussed.

After motivating and defining the challenge problem and the SPLID dataset in Sections 2-4, Section 5 discusses the baseline analytical algorithmic and machine learning PoL node detection algorithms. Finally, Section 6 concludes the paper and provides recommendations for further improvements.

2. PROBLEM DEFINITION

GEO satellite station-keeping routines are dictated by a combination of human-driven variables, hardware limitations, and complex physics, all of which introduce inconsistencies that make it difficult or impossible for human analysts to characterize the features and patterns of nominal maneuvers. For example, holiday schedules, thruster degradation, and fluctuating perturbations can influence the frequency, magnitude, and direction of a satellite's station-keeping maneuvers to evolve over time. Figure 1 shows a six-month geodetic coordinate history for *ArabSat 5A* during which the satellite performed consistent north-south station-keeping maneuvers every 14 days but irregular east-west maneuvers ranging from 11 to 18 days apart. Discrepancies like this are often indiscernible to human analysts but are increasingly prevalent as satellites age. Characterizing behavior patterns for groups of similarly-classed satellites is no less

[1] `https://github.com/ARCLab-MIT/SPLID_Benchmark_Suite`

challenging because there is a lack of common operational guidelines for station-keeping protocols among satellite operators. These protocols can differ even for satellites with the same satellite bus and operated by the same operator.

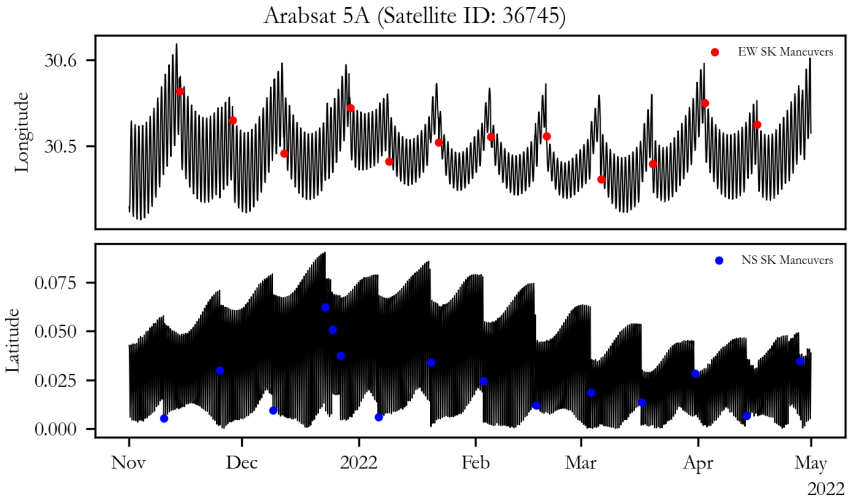

Fig. 1: TLE-derived geographic data for *ArabSat 5A* over a 6-month period during a station-keeping behavioral mode. Although the satellite is operating with the same behavioral mode, variation in control strategies can be observed. Figure adapted from [15].

In recent years, an increasing number of satellite manufacturers have moved toward modular busses that support multiple propulsion options, resulting in satellites requisitioned by the same operator, at the same time, with the same reported bus having different propulsion capabilities. This can have a significant impact on the satellites' station-keeping routines since different types of propulsion provide distinct advantages and considerations.

Mobile GEO satellites use chemical and/or electric propulsion systems to perform station-keeping maneuvers. Chemical propulsion systems use combustion engines to generate thrust via the combustion of propellant. They can generate a higher thrust-to-weight ratio compared to electric propulsion systems and are often characterized by the large delta-V (ΔV) required to transition between adjacent satellite states. On the other hand, electric propulsion systems use electrical and/or magnetic force to accelerate a propellant as a means of generating thrust. Electric propulsion systems require comparatively little mass to accelerate a spacecraft and are generally used for small corrections in the spacecraft's attitude due to their smaller thrust output. Electric propulsion systems also allow for finer control of their thrust output and extended-duration burns. Some satellites station-keep using both chemical and electric propulsion, or they use electrothermal thrusters which exhibit a mixture of chemical and electric propulsion characteristics. This work refers to such satellites as having hybrid propulsion systems.

Satellites with both chemical and electric thrusters may change which system they use for station-keeping throughout their operational lifetimes. Several factors influence this choice, including the satellite's mission objectives, fuel reserves, and the health of its propulsion system. Towards the end of the operational lifetime, satellite operators might consider restricting the usage of chemical propulsion for station-keeping to conserve fuel and prolong the satellite's longevity. Equally significant is the health of the propulsion system as degraded thrusters or lost fuel can necessitate the adaptation of a different control strategy. For all of these reasons, propulsion capabilities and usage are relevant features of a satellite's station-keeping behavior and PoL. However, it is time-consuming, challenging, or even impossible to identify the correct station-keeping propulsion type from visual inspection of astrometric data alone.

To that end, the satellite Pattern-of-Life (PoL) characterization problem focuses on using AI algorithms to develop an automated framework that can efficiently characterize and detect changes in the behavioral patterns of GEO satellites using multi-faceted SSA data. Specific features of interest include periods of station-keeping and the type of propulsion—chemical, electric, or hybrid—used to perform nominal station-keeping maneuvers. A satellite's PoL consists of two fundamental components: nodes and behavioral modes. A node represents an instantaneous point on the PoL timeline that separates two distinct behavioral modes. For this challenge problem, participants are tasked to write an AI algorithm to detect the nodes and classify the behavioral modes for a number of GEO satellite PoLs span-

ning a six-month study period and to organize their results in a standard format specified in Section 3. In this case, there are four behavioral mode classes—not station-keeping (NK), station-keeping with electric propulsion (EK), station-keeping with chemical propulsion (CK), and station-keeping with hybrid propulsion (HK)—for both the longitudinal (east-west) and transverse (north-south) directions. The corresponding PoL nodes are therefore identified by one of three labels—initiate drift (ID), adjust drift (AD), and initiate station-keeping (IK).

An ID node occurs when a satellite that has previously been station-keeping leaves that station and starts drifting in longitudinal space. AD nodes then capture any significant changes in longitude drift rate or direction for a drifting satellite. An IK node indicates a change in a satellite's behavioral mode from drifting to station-keeping—a behavior pattern in which the satellite's orbital position is maintained within a small range of latitudes and longitudes. These ranges can vary between satellites with different ages, propulsion systems, operators, and mission objectives. An example of this variation is shown in Figure 2 which compares TLE-derived longitude histories for two satellites—-*ASTRA 1f* which launched in 1996 with a chemical-only propulsion system and *Eutelsat 172B* which was launched in 2017 and equipped with an electric-only propulsion system. Both satellites conduct north-south and east-west station-keeping during the two-month period, but while *ASTRA 1F* performs independent north-south and east-west maneuvers, *Eutelsat 172B* performs composite station-keeping.

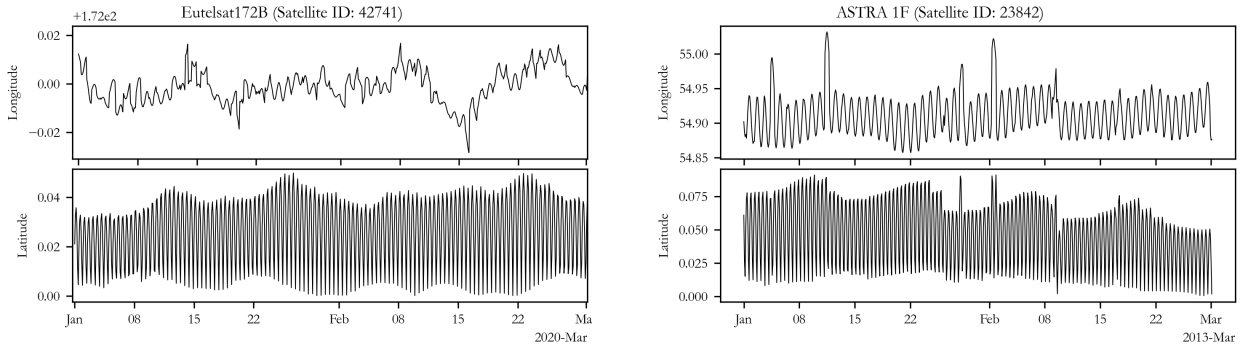

Fig. 2: TLE-derived latitude and longitude histories for two commercial geostationary communications satellites operating in station-keeping behavioral modes over a two month period. (a) *Eutelsat 172B* (Satellite ID: 42741) has used electric thrusters to perform station-keeping maneuvers since it became active in 2017 and maintains a station size of 0.05 degrees longitude and 0.05 degrees latitude during this PoLbehavioral mode. (b) *ASTRA 1f* (Satellite ID: 23842) used chemical thrusters to perform station-keeping maneuvers prior to its retirement in 2020 and maintained a station size of 0.1 degrees longitude and 0.08 degrees latitude during this PoL behavioral mode. Figures taken from [15].

When conducting north-south and east-west station-keeping independently, a satellite performs out-of-plane maneuvers to ensure that its latitudinal position remains within the latitude deadband and additional in-plane maneuvers to stay within its longitude deadband. In contrast, composite station-keeping is a method of combining north-south and east-west maneuvers and is commonly employed by satellites with electric propulsion capabilities and especially those with gridded ion systems due to their smaller thrust and higher precision compared to chemical propulsion systems. In general, maneuvers are much more frequent in composite station-keeping routines than in routines with independent north-south and east-west components. This is indeed the case for the station-keeping routines exhibited in Figure 2, and the effect is that *Eutelsat 172B* achieves a smaller deadband compared to *ASTRA 1f*. PoL nodes and behavioral modes may be analyzed according to the same convention in either case by using independent longitudinal and transverse labels. PoLs with composite station-keeping will simply have identical labels and timestamps in the north-south and east-west directions.

Figure 3 contains additional examples of composite and independent station-keeping in the longitudinal histories of *Horizons 3E* and *Galaxy 17*, for which the individual station-keeping maneuvers are annotated. *Horizons 3E* uses the electric Xenon Ion Propulsion System for composite station-keeping twice daily. On the other hand, *Galaxy 17* uses a chemical propulsion system to carry out weekly east-west maneuvers and separate north-south station-keeping every other week. Note that *Horizons 3E* completes multiple station-keeping cycles during its nine-day study period while a study period of at least a month is required to see two cycles of *Galaxy 17*'s station-keeping routine. This demonstrates that some station-keeping protocols must be characterized from longer time histories than others, and that the study

period chosen for the participants constrains observable station-keeping routines to those with at least two cycles within six months. To establish a convention in the context of the challenge problem, a satellite may be considered as not station-keeping (NK) in either the longitudinal (EW) or transverse (NS) direction if it does not performed a station-keeping maneuver in that direction at a frequency less than once every 60 days. Furthermore, transverse station-keeping behavioral modes should not coincide with NK longitudinal modes. This is because the provided data set consists of low-inclination, geostationary satellites that occupy orbital stations at consistent longitudes, and though inclination corrections may occur outside of such stations, they should not be interpreted as station-keeping behavior in this context.

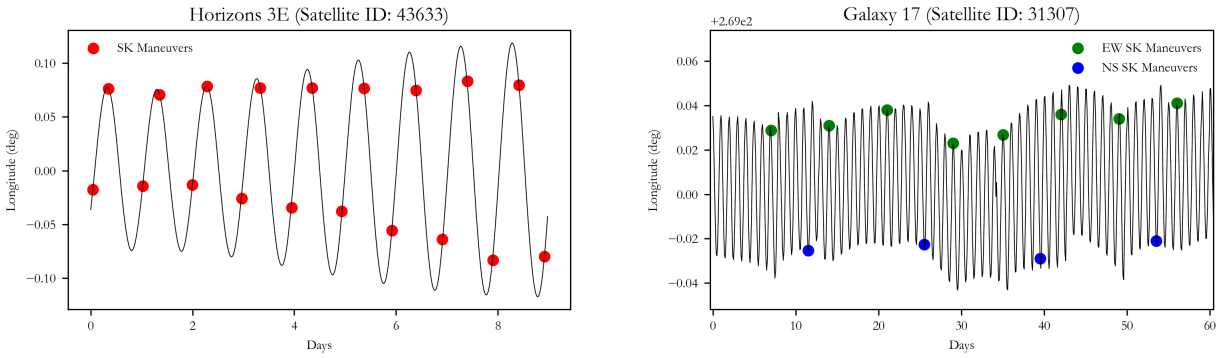

Fig. 3: Longitude histories for two satellites operating in station-keeping behavioral modes with different maneuver routines. (a) *Horizons 3E* (Satellite ID: 43633) performs twice-daily composite station-keeping (SK) maneuvers (marked in red) using an electric propulsion system. 19 maneuvers were performed over this 9-day history derived from Intelsat positional data [16]. (b) *Galaxy 17* (Satellite ID: 31307) performs weekly east-west SK maneuvers (green) and biweekly north-south SK maneuvers (blue) using a chemical propulsion system. 12 maneuvers were performed over this 60-day history derived from VCMs.

Other notable conventions involve transverse and longitudinal node placement. Figure 4 provides a visual representation of the PoL nodes associated with the satellite *Horizons 2* over a span of two and a half months, from mid-September through November of 2022. These PoL nodes were annotated by a subject matter expert with the help of external metadata, such as satellite bus type and manufacturer information. Over this time frame, the satellite exhibits several shifts in its behavioral mode, each of which is distinctly marked by a PoL node. The figure illustrates a conventional repositioning operation, where a GEO satellite performs a sequence of deliberate maneuvers that shift it to a new longitude position. The satellite was initially drifting at the start of the study period and conducted two longitudinal drift adjustments before transitioning into an east-west station-keeping behavioral mode. With the assistance of external metadata, the satellite was identified to be using an electrothermal (hybrid) propulsion system for its east-west station-keeping maneuvers. Although the satellite was performing inclination corrections during its period of longitudinal drift, it did not achieve the nominal inclination for its new station until mid-October, at which time it had been performing east-west station-keeping for nearly two weeks. Therefore, the north-south station-keeping node is assigned to the timestamp of the maneuver that lowered the satellite's inclination into the nominal range.

While latitude, longitude, and inclination histories can provide visualizations of station-keeping maneuvers and PoL characteristics like those in Figures 3 and 4, they may obscure the limitations of the data from which they are derived. This is also true for orbital element sets and other histories reconstructed from another pre-processed data source, but utilizing multiple coordinate systems can enable more holistic analyses of satellite behavior. Transforming large astrometric data sets between coordinate and reference systems can be overwhelming to those unfamiliar with the astrophysical concepts involved. However, to remove this barrier to entry, participants will be provided an interpolated data set containing multiple coordinate and state formats.

3. DATASET

This section provides a high-level overview of SPLID. SPLID comprises a public challenge dataset and a private evaluation dataset. The public challenge dataset consists of 500 simulated satellite trajectories, each operating under different mission objectives and equipped with different propulsion capabilities. Meanwhile, the private evaluation

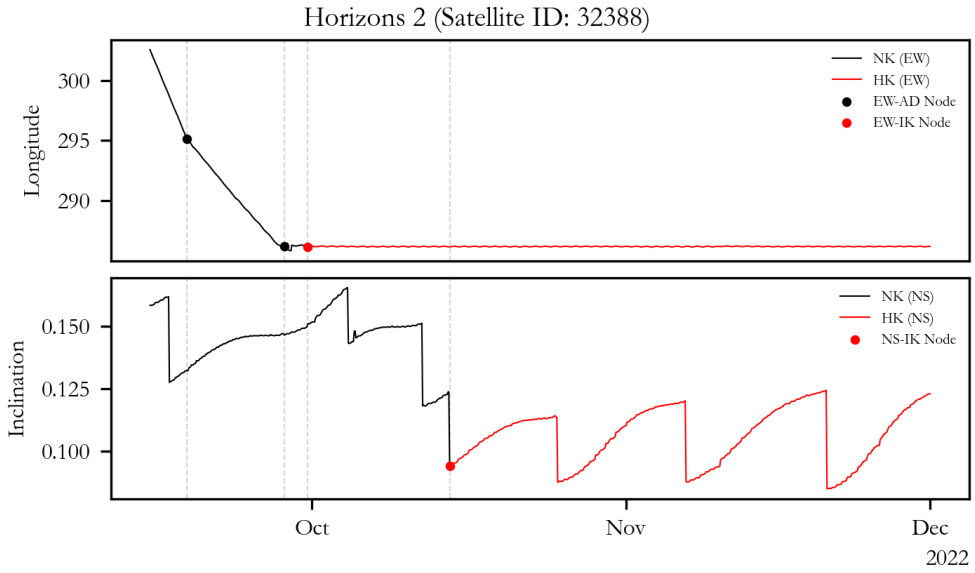

Fig. 4: VCM-derived longitudinal and inclination histories for *Horizons 2*, annotated by longitudinal and transverse PoL components respectively. Trace colors indicate the behavioral mode as either station-keeping with hybrid propulsion (red) or not station-keeping (black). Dashed vertical lines indicate the timestamp of each PoL node within both signals while circular markers denote directional PoL nodes within their respective domains. Nomenclatures for the PoL nodes and behavioral modes are provided in Table 1.

dataset consists of a mixture of 100 simulated satellite trajectories, and 125 historical satellite trajectories generated from VCM data and high-accuracy ephemerides provided by satellite operators. The astrometric data consists of the osculating orbital elements, geodetic positions, and the Cartesian position and velocity in the J2000 inertial reference frame over six months at a two-hour temporal resolution. Figure 5 shows the entries for one of the satellites within the dataset.

	Eccentricity	Semimajor axis (km)	Inclination (deg)	RAAN (deg)	Argument of periapsis (deg)	True anomaly (deg)	Latitude (deg)	Longitude (deg)	Altitude (km)	J2k X (km)	J2k Y (km)	J2k Z (km)	J2k Vx (km/s)	J2k Vy (km/s)	J2k Vz (km/s)
0	0.000345	42322.787327	0.134197	96.741868	4.601943	-150.663032	0.005163	330.828873	35957.361552	27596.077884	-32105.261631	-55.360166	2.326261	2.000231	-0.005961
1	0.000342	42323.635493	0.134172	96.727355	0.271599	-116.419371	-0.003316	330.645541	35951.937964	39921.569096	-14074.644427	-88.981686	1.019342	2.894129	-0.003165
2	0.000334	42324.053073	0.134196	96.619165	357.885514	-84.016686	-0.010939	330.471816	35944.440373	41615.381839	7703.942803	-98.900236	-0.559640	3.017496	0.000487
3	0.000351	42324.053596	0.134197	96.617227	358.997513	-55.207751	-0.015547	330.308679	35937.427673	32221.481778	27429.478796	-82.369402	-1.990335	2.336697	0.004000
4	0.000339	42323.084988	0.134415	96.531735	2.682806	-28.880005	-0.015959	330.154093	35932.384820	14238.611758	39842.693657	-43.819258	-2.890905	1.032598	0.006462
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2168	0.000273	42165.897813	0.135902	95.875192	243.819284	-15.582876	0.001133	286.192416	35776.669780	34152.064260	-24711.508360	-74.581104	1.802640	2.491704	-0.004858
2169	0.000249	42165.339621	0.136829	96.014433	247.379333	10.813192	-0.006927	286.205639	35776.902629	41939.635327	-4254.951325	-98.540632	0.310555	3.059648	-0.001503
2170	0.000241	42165.339461	0.136829	96.012254	251.515960	36.773045	-0.012669	286.217805	35779.065333	38421.717362	17349.125979	-95.590548	-1.265148	2.802912	0.002304
2171	0.000232	42165.754868	0.137116	95.893065	256.724823	61.773288	-0.014994	286.225589	35782.987772	24548.158706	34277.453849	-66.858540	-2.499588	1.790890	0.005510
2172	0.000213	42165.754523	0.137118	95.892590	257.906890	90.675389	-0.013118	286.227148	35787.722053	4059.803848	41969.956274	-19.976065	-3.060239	0.296683	0.007212

Fig. 5: Astrometric data (osculating orbital element, geodetic position, and the Cartesian position and velocity in the J2000 inertial reference frame) for a satellite over a six-month period at a two-hour temporal resolution.

Each data is accompanied by a list of expert-annotated time-stamped pattern-of-life nodes. Figure 6 shows the time-stamped satellite PoL nodes for one of the satellites within the data set. Each row in the satellite PoL node list begins with a time index of the node, followed by the direction (EW - east-west, NS - north-south), the node type, and the propulsion type observed in the following behavioral mode. The time index refers to the index (row) of the astrometric data that corresponds to that particular PoL node. The descriptions of each

	Time Index	Direction	Node	Type
0	0	EW	SS	NK
1	0	NS	SS	NK
2	212	EW	AD	NK
3	327	EW	AD	NK
4	355	EW	IK	CK
5	523	NS	IK	CK
6	2172	ES	ES	ES

Fig. 6: Time-stamped satellite PoL nodes.

label are provided in Table 1.

Table 1: Descriptions for satellite behavioral mode labels.

Node Label	Description	Type Label	Description
SS	Start of the study period	NK	Not station-keeping
ES	End of the study period	CK	Station-keeping using chemical propulsion system
ID	Initiate drift	EK	Station-keeping using electric propulsion system
AD	Adjust drift	HK	Station-keeping using hybrid propulsion system
IK	Initiate station-keeping		

3.1 Creation of the synthetic dataset

An in-house satellite simulation tool developed by the MIT Lincoln Laboratory is used to generate the synthetic dataset. The simulation tool uses a high-fidelity satellite propagator [17] to simulate the trajectory of the satellite under different operating objectives and propulsion capabilities. The high-fidelity satellite propagator uses the special perturbations approach to allow for the precise modeling of the satellite's perturbed motion. It explicitly solves for the perturbation forces that are acting on the satellite at any point in space and time. Perturbations due to the Earth's non-uniform gravitational field, gravitational forces of third bodies (the moon and sun), atmospheric drag forces, solar radiation pressure, as well as perturbation effects caused by solid Earth tides, ocean tides, and general relativity are taken in consideration when generating the synthetic dataset.

Satellites with varying physical parameters and orbital elements are simulated to operate with time-varying operating objectives over six months. Satellites can freely change their behavioral modes independently of each other. In this case, behavioral modes are broadly classified as either drifting in longitudinal space or station-keeping and by the type of propulsion the satellite is using to maneuver. This class scheme allows for five possible labels for each direction—initiate drift, adjust drift, initiate station-keeping protocol using a chemical propulsion system, initiate station-keeping protocol using an electric propulsion system, and initiate station-keeping protocol using a hybrid (combination of chemical and electric) propulsion system.

3.2 Creation of the VCM dataset

RSO ephemerides from a high-fidelity special perturbations orbit propagator with assimilated tracking observations are issued by the U.S. JSpOC in the form of VCM data. Each VCM includes the position and velocity vector of the RSO expressed in Earth-Centered Inertial and Earth-centered Earth-fixed reference frames, the covariance of the estimate, and additional metadata related to the orbital perturbation model used by the orbit propagator. The VCM data are reported at non-uniform time intervals, hence a special perturbation propagator is used to generate the astrometric data with evenly spaced time steps to be used for the challenge problem. The special perturbation propagator is developed using the Python Poliastro package [18]. RSO states are propagated using the Cowell's formulation to a two-hour temporal resolution, with the inclusion of the J2 and third bodies perturbations.

The VCM astrometric data are then manually annotated by a human expert to generate a list of time-stamped satellite PoL nodes. Here, external metadata, such as a satellite's bus type and age, is used to assist in labeling the correct propulsion type for the station-keeping nodes.

3.3 Collection of owner-operator dataset

Operational satellites are routinely tracked by their owners and operators and some of these data are made publicly available. CelesTrak curates a list of supplemental General Perturbations element sets derived directly from owner-operator supplied orbital data [19]. These owner-operator data are more accurate than the two-line element set (TLE) released by the U.S. JSpOC [19]. Furthermore, these data are released routinely in a timely manner as part of their normal operations. Some satellite owner-operators, such as Intelsat [16], also provide additional data on when station-keeping and relocation maneuvers are scheduled and can be used to assist in accurately identifying the satellite's behavioral mode.

Similarly to the VCM dataset, the owner-operator data are propagated to a two-hour temporal resolution and then manually annotated by a human expert to generate the list of time-stamped satellite behavioral modes.

4. COMPETITION DESIGN

In order to lower the barrier of entry, a development kit is provided to entrants in addition to the SPLID dataset. The development kit is coded in both Matlab and Python and consists of a set of basic utility functions and tutorials to help the participants get started with the challenge problem. The tutorials will guide the participants in reading and understanding the data, parsing and manipulating the data, training and evaluating the baseline ML algorithm, and submitting their ML algorithms to the competition platform.

The following subsections outline the design of the AI SSA challenge problem, where the competition setup and evaluation metric are defined.

4.1 Competition Setup

The competition is hosted using the open-source EvalAI platform. The EvalAI platform provides automation for key challenge problem features, such as participant registration and submission, implementation of evaluation metrics, live leaderboard, and forum for participant discussions. One of the main advantages of using the EvalAI platform is that it allows for code upload-based challenges. Instead of having participants submit their model predictions on a static test set, the participants will be submitting their trained algorithm with their training code/workflow to the competition platform, in the programming language they prefer[2]. The performance of their trained AI algorithm will then be evaluated against our internal test data and evaluation metric on our remote server. Once the evaluation is complete, the results are sent back to the live leaderboard on the challenge platform. This allows for the usage of a private test set, which helps with data privacy concerns and prevents data leakage.

During the competition stage, only 25% of the test set will be used for evaluation and placement in the public leaderboard. In addition to the public leaderboard, there will also be a private internal leaderboard that tracks the model performance using the full test set. The performance on the private leaderboard will be released at the end of the competition and the final ranking will be determined based on their standing in the private leaderboard. This is done to prevent the participants from overfitting their AI algorithms to the test data. In the event of an exact score tie, the tiebreaker will be the algorithms' submission dates, with an advantage given to participants who submit early.

Although the participants are only provided with synthetic data for training, the trained model will also be evaluated on real data based on JSpOC-published VCMs and owner-operator ephemerides. Each has different data qualities, in terms of error, uncertainties, and temporal resolution. This will provide insight into how transferable the trained agent to real-life applications is.

4.2 Evaluation Metrics

The performance of the participant submissions will be evaluated based on their ability to accurately detect and classify satellite behavioral mode changes compared to the ground truth labels. The key metric used to create the leaderboard is the F_2 **score**, which is a variant of the F1 score that emphasizes recall over precision. The precision corresponds to the proportion of detected nodes (True Positives (TPs) and False Positives (FPs)) that were correct, whereas the recall corresponds to the percentage of actual true nodes (True Positives (TPs) and False Negatives (FNs)) that were detected. Given that, the F_2 score is given as:

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}, \quad F_2 = \frac{5(\text{Precision} \times \text{Recall})}{4\text{Precision} + \text{Recall}} \tag{1}$$

To determine whether a node detection is correct or not, i.e., to define what TPs, FPs and FNs mean in this context, we consider two main criteria:

1. Data mislabelling.
2. Mistiming of the label, controlled by a tolerance of 6 time indices (12 hours before and after the node).

Based on these criteria, the definition of TPs, FPs, and FNs is given in Table 2. Besides, Figure 7 shows an example of how to evaluate a participant submission against the ground truth for one object of the dataset. The propulsion type

[2]Running environments will be provided as containers for common languages, but participants will be allowed to submit with other custom containers too

Table 2: Definition of true positive (TP), false positive (FP), and false negative (FN)

Acronym	Description
TP	The participant node falls within the time tolerance interval of the ground truth node, and the labels match
FP	Two possibilities: (1) The participant node falls within the time tolerance interval of a ground truth node, but the labels do not match. (2) There are no ground truth nodes within the tolerance interval around the participant node.
FN	Missed ground truth node; there is no participant node within the tolerance interval of a ground truth node

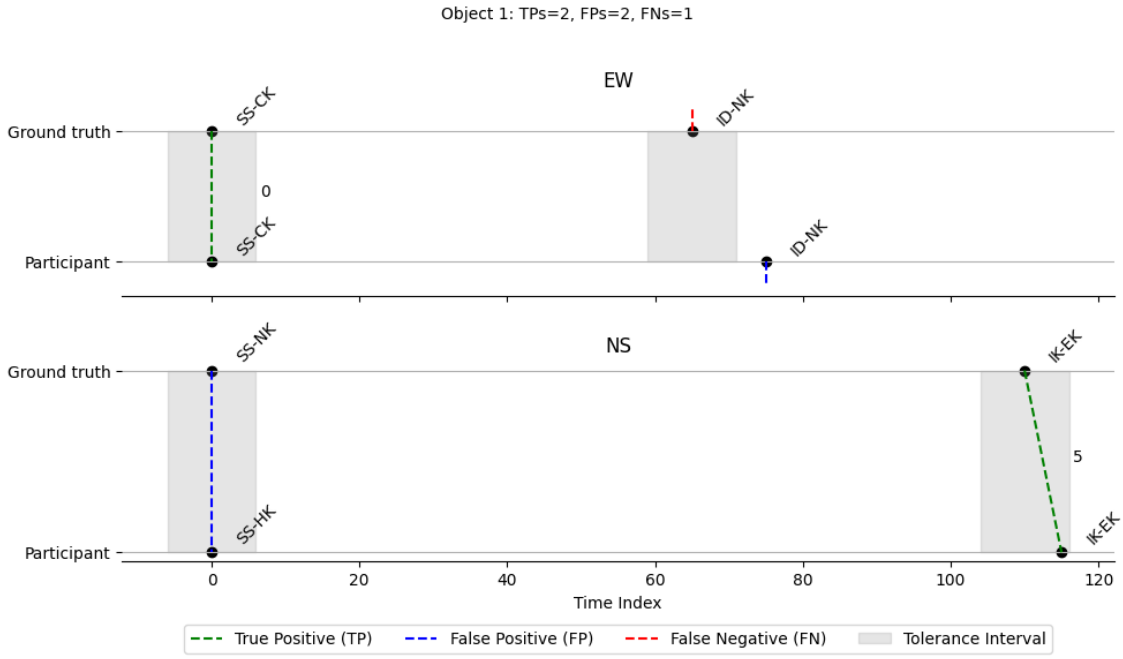

Fig. 7: Example of the evaluation process for one object. Subplots show 'EW' and 'NS' directions, with ground truth (upper axis) and participant-detected nodes (lower axis). Numbers next to the True Positive predictions represent the distance, in time indices, between the participant node and the corresponding ground truth node.

and name of the node conform the label of the node (e.g. SS-NK). The evaluation is run for every node present in the ground truth, both in EW and NS. For each of them, we look for the first participant node within a certain tolerance interval around the time index of the node. If the labels of the participant and ground truth nodes match, that sums up a true positive (see the first EW ground truth node in the figure). Otherwise, we sum up a false positive (see the first NS ground truth node in the figure). In case no participant node falls within the interval around the ground truth node, we sum up a false negative (see the second EW ground truth node). Finally, participant nodes that are not matched to any ground truth node (e.g., node ID-NK direction EW in Fig. 7 are considered as false positives too.

Even if two participants have the same classification results, and therefore, the same F_2 score, the final score will also take into account a component that will penalize the mistiming of correct assignments, so that two correctly predicted nodes can contribute differently to the final metric based on the distance between the predicted node and the actual node (See, as an example, the differences between the two TPs shown in Fig. 7). More details about this and the final formulation of the score, will be given in a specific section of the challenge website once it is released.

5. BASELINE SOLUTIONS

Two baseline solutions have been set up and will be provided along with a development kit with the SPLID dataset. The first baseline solution uses a heuristic-based approach. This approach builds upon the algorithmic PoL characterization model introduced in [2]. In this extension, the algorithmic PoL characterization model further categorizes the station-keeping nodes based on the specific propulsion system used. On the other hand, the second baseline solution is a machine learning-based approach that uses random forests with lagged features for satellite PoL identification.

These baselines will help facilitate the assessment of more advanced AI solutions and help familiarize the participants with the dataset, methodologies, and evaluation pipeline. Participants who are new to the field can leverage the baseline implementation as a starting point and build upon the baseline implementation by iterating on the existing model and experimenting with modifications and incorporating newer AI techniques. This will lower the barrier of entry and encourages broader and more diverse participation.

5.1 Heuristic-based Approach

The heuristic-based approach detects changes in the satellite's longitude and inclination waveform by analyzing the variation in their standard deviation. The localized standard deviation is calculated for each data point over the last orbital period—around 24 hours for GEO satellites. Subsequently, the algorithmic PoL characterization model examines each data point over the six-month period and identifies instances where the standard deviation is greater than a threshold associated with the most common varieties of station-keeping: the maximum value typically observed within station-keeping behavioral modes. The first occurrence of such an event indicates a potential initiate drift (ID) node, suggesting that the satellite has transitioned from a station-keeping behavioral mode to a drifting behavioral mode. During the period of elevated standard deviation, an adjust drift (AD) node is assigned if the variation in consecutive standard deviation values is greater than 10%. This would indicate that the satellite's drift rate or direction has changed during the longitudinal shift maneuver.

When the standard deviation falls back below the stationary mode threshold, this would indicate that the satellites might have transitioned back to a station-keeping behavioral mode. During the low standard deviation period, the oscillation frequency and amplitude of the waveform are analyzed to determine the propulsion type utilized in the station-keeping behavioral modes. In the study conducted by Solera et al. (2023), the heuristic-based approach demonstrated notable performance by correctly labeling approximately 83% of the PoL nodes, 369 out of 444 instances, with an additional 248 false positive. Interestingly, many of these false positive labels were time-stamped with the correct epoch but was assigned an incorrect node label as per the ground truth data [2]. In their analysis, the heuristic-based approach was

Table 3: Performance of the heuristic-based approach

Node Label	Number of Samples	True Positive	False Positive	Precision	Recall	F_1 score
ID	125	106	7	0.9381	0.8480	0.8908
AD	189	154	239	0.3919	0.8148	0.5292
IK	130	109	2	0.9820	0.8385	0.9046

scored solely on its capacity to accurately predict the node label, while the type label was factored into the evaluation. The precision, recall, and F1 score of the heuristic-based approach is shown in Table 3. The heuristic-based approach has an average F_1 score of 0.7748. Readers are referred to [2] for more information on the heuristic-based approach. While it successfully identified a large portion of the PoL nodes, it struggles with accurately identifying AD nodes and there is a need to further enhance its performance. It is important to note, however, that the approach was only evaluated on the node label and the performance are expected to further drop when the type label is incorporated into the analysis. The addition of the node type is anticipated to introduce additional complexity which can be challenging for the heuristic-based approach.

5.2 Machine Learning-based Approach

A random forest classifier model [20] was developed as a baseline machine learning approach for the challenge. The model takes in the multi-dimensional satellite state data and predicts the PoL node labels at each time step. To enable classic tabular machine learning, the model has to be trained to predict labels at each time step (row in the table), but its

performance must be evaluated based on change point detection accuracy compared to the ground truth. Additionally, since there can be more than one node at the same time step (one for EW and one for NS), separate models have to be trained for each direction.

ObjectID	Time Index	Feat. 1	Feat. 2	...
36391	0	50	15	...
36391	1	40	25	...
36391	2	30	35	...
...	...	...	...	...
43401	0	20	45	...
43401	1	10	55	...
...	...	...	...	...

(a) Input data

ObjectID	Time Index	Direction	Node	Type
36131	0	EW	SS	CK
36131	25	EW	ID	CK
36131	0	NS	SS	NK
36131	110	NS	IK	EK
43401	0	EW	SS	HK
43401	0	NS	SS	EK
...	...	...	...	...

(b) Ground truth data

ObjectID	Time Index	Feat. 1	Feat. 1 (lag 1)	Feat. 1 (lag 2)	Feat. 2	...	EW	NS
36391	0	50	NaN	NaN	15	...	SS-CK	SS-NK
36391	1	40	50	NaN	25	...	NaN	NaN
36391	2	30	40	50	35	...	NaN	NaN
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
43401	0	20	NaN	NaN	45	...	SS-HK	SS-EK
43401	1	10	20	NaN	55	...	NaN	NaN
...	...	...	...	...	...	...	...	...

(c) Combined input and ground truth, with lagged features, and EW and NS as target columns

ObjectID	Time Index	Feat. 1	Feat. 1 (lag 1)	Feat. 1 (lag 2)	Feat. 2	...	EW	NS
36391	0	50	50	50	15	...	SS-CK	SS-NK
36391	1	40	50	50	25	...	SS-CK	SS-NK
36391	2	30	40	50	35	...	SS-CK	SS-NK
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
43401	0	20	20	20	45	...	SS-HK	SS-EK
43401	1	10	20	20	55	...	SS-HK	SS-EK
...	...	...	...	...	...	...	...	...

(d) Final dataframe ready for tabular machine learning, without missing values

Fig. 8: Data transformation process for preparing the data for tabular machine learning.

To meet these requirements, the ground truth data is first separated into EW and NS types. The "node" and "type" columns are concatenated to create a single label for both directions EW and NS, such as 'SS-CK'. These labels are forward filled so each row has a label. Additionally, the input data is expanded by creating lagged features from the raw feature columns (position, velocity, orbital elements, ...). Lagged features capture historical context by using the prior n time steps of each feature. The final data table with labels is then split into training and validation sets by satellite *ObjectID* to avoid data leakage. A representation of the data transformation process is shown in Fig. 8.

The model predictions are post-processed to convert the per time step predictions into a format matching the ground truth (See Fig. 6). To do this, the predicted EW and NS labels are first split back into the original "node" and "type" columns. Then, they are filtered to only keep rows where the label changes compared to the prior time step. This process converts both the ground truth and predictions into a consistent format with labels provided only at change points.

The approach is evaluated using the evaluation metrics of precision, recall, and F_1 score (see Section 4.2) on a toy

dataset of 49 objects for training+validation, and 22 objects for test. While the described approach achieves good scores on the training set (Precision=0.69; Recall=1.0; F_1=0.82), the performance drops in the validation set (Precision: 0.07 Recall: 0.01 F_1 Score: 0.02) and in the test set (Precision: 0.53 Recall: 0.01 F_1 Score: 0.02). Although this is clearly a sign of overfitting due to the usage of just a small portion of the data, it also hints at the difficulty of the proposed challenge, and how more complex models are needed. The high dimensionality of the problem and the presence of multiple interleaved time series make this problem unique for machine learning and requires of representations and techniques that, unlike the above baseline, capture temporal relationships in an accurate and generalisable way.

6. CONCLUSION

As the near-Earth space environment becomes increasingly congested, it becomes ever more important to develop more efficient RSO tracking and orbit prediction capabilities. One efficient approach to achieve this is by accurately characterizing the satellite's behavioral mode. To facilitate the development of advanced AI algorithms for automated satellite pattern-of-life characterization using astrometric time-series data, the open-source SPLID dataset is curated. In conjunction with the release of the SPLID dataset, an AI SSA challenge competition is hosted. This competition aims to foster innovative AI applications in the domain of Space Situational Awareness while also providing valuable insights to the SSA research community regarding the capabilities and limitations of AI algorithms. Two baseline solutions are provided along with the SPLID dataset to assist the participants in developing their AI model.

ACKNOWLEDGEMENT

Research was sponsored by the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 2141064. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Daniel Jang is supported by the MIT Lincoln Laboratory Scholarship Program Fellowship. The authors also thank the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC, database, and consultation resources that have contributed to the research results reported in this paper.

REFERENCES

[1] Thomas G Roberts, Haley E Solera, and Richard Linares. Geosynchronous satellite behavior classification via unsupervised machine learning. In *9th Space Traffic Management Conference, Austin, TX*, 2023.

[2] Haley E Solera, Thomas G Roberts, and Richard Linares. Geosynchronous satellite pattern of life node detection and classification. In *9th Space Traffic Management Conference, Austin, TX*, 2023.

[3] T. G. Roberts and R. Linares. A Survey of Longitudinal-Shift Maneuvers Performed by Geosynchronous Satellites from 2010 to 2021. In *73rd International Astronautical Congress*, 2022.

[4] P. DiBona, J. Foster, A. Falcone, and M. Czajkowski. Machine learning for RSO maneuver classification and orbital pattern prediction. In *20th Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS)*, Maui, HI, 2019.

[5] Liang Liu, Ling Tian, Zhao Kang, and Tianqi Wan. Spacecraft anomaly detection with attention temporal convolution networks. *Neural Computing and Applications*, pages 1–9, 2023.

[6] Francisco Caldas and Cláudia Soares. Machine learning in orbit estimation: a survey, 2023.

[7] John C Dorelli, Chris Bard, Thomas Y Chen, Daniel Da Silva, Luiz Fernando Guides dos Santos, Jack Ireland, Michael Kirk, Ryan McGranaghan, Ayris Narock, Teresa Nieves-Chinchilla, et al. Deep learning for space weather prediction: Bridging the gap between heliophysics data and theory. *arXiv preprint arXiv:2212.13328*, 2022.

[8] Weiwei Jiang. Time series classification: Nearest neighbor versus deep learning models. *SN Applied Sciences*, 2(4):721, 2020.

[9] Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

[10] Mate Kisantal, Sumant Sharma, Tae Ha Park, Dario Izzo, Marcus Märtens, and Simone D'Amico. Satellite pose estimation challenge: Dataset, competition design, and results. *IEEE Transactions on Aerospace and Electronic Systems*, 56(5):4083–4098, 2020.

[11] Bo Chen, Daqi Liu, Tat-Jun Chin, Mark Rutten, Dawa Derksen, Marcus Martens, Moritz von Looz, Gurvan Lecuyer, and Dario Izzo. Spot the geo satellites: From dataset to kelvins spotgeo challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2086–2094, 2021.

[12] Thomas Uriot, Dario Izzo, Luís F Simões, Rasit Abay, Nils Einecke, Sven Rebhan, Jose Martinez-Heras, Francesca Letizia, Jan Siminski, and Klaus Merz. Spacecraft collision avoidance challenge: Design and results of a machine learning competition. *Astrodynamics*, 6(2):121–140, 2022.

[13] Tae Ha Park, Marcus Märtens, Mohsi Jawaid, Zi Wang, Bo Chen, Tat-Jun Chin, Dario Izzo, and Simone D'Amico. Satellite pose estimation competition 2021: Results and analyses. *Acta Astronautica*, 204:640–665, 2023.

[14] Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shiv Baran Singh, Stefan Lee, and Dhruv Batra. Evalai: Towards better evaluation of ai agents.

[15] Haley Elizabeth Solera. Python-based tools for characterizing geosynchronous satellite behavior and evaluating maneuver prediction techniques. Master's thesis, Massachusetts Institute of Technology, 2023.

[16] Intelsat. Intelsat ephemeris data. `https://my.intelsat.com/ephemeris/public#/`. Retrieved July 30, 2023.

[17] Meysam Mahooti. High precision orbit propagator. `https://www.mathworks.com/matlabcentral/fileexchange/55167-high-precision-orbit-propagator`, MATLAB Central File Exchange. Retrieved July 30, 2023.

[18] Juan Luis Cano Rodríguez, Yash Gondhalekar, Antonio Hidalgo, Shreyas Bapat, Nikita Astrakhantsev, Chatziargyriou Eleftheria, Kevin Charls, Meu, Dani, Abhishek Chaurasia, Alberto Lorenzo Márquez, Dhruv Sondhi, Tomek Mrugalski, Emily Selwood, Manuel López-Ibáñez, Orestis Ousoultzoglou, Pablo Rodríguez Robles, Greg Lindahl, Syed Osama Hussain, andrea carballo, Andrej Rode, Helge Eichhorn, Anish, sme, Himanshu Garg, Hrishikesh Goyal, Ian DesJardin, Matthew Feickert, and Ole Streicher. poliastro/poliastro: poliastro 0.17.0 (SciPy US '22 edition), July 2022.

[19] T.S. Kelso. Supplemental general perturbation element sets. `https://celestrak.org/NORAD/elements/supplemental/`. Retrieved July 30, 2023.

[20] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.