Received 30 January 2025, accepted 11 February 2025, date of publication 19 February 2025, date of current version 28 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3543813

RESEARCH ARTICLE

Trade-Space Exploration With Data Preprocessing and Machine Learning for Satellite Anomalies Reliability Classification

ABDUL MUTHOLIB[1,2], (Student Member, IEEE), NADIRAH ABDUL RAHIM[1], (Member, IEEE), TEDDY SURYA GUNAWAN[1], (Senior Member, IEEE), AND MIRA KARTIWI[3], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, Faculty of Engineering, International Islamic University Malaysia, Kuala Lumpur 53100, Malaysia
[2]Department of Information System, Faculty of Science and Technology, UIN Syarif Hidayatullah Jakarta, Banten 15412, Indonesia
[3]Department of Informations Systems, Faculty of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur 53100, Malaysia

Corresponding author: Nadirah Abdul Rahim (nadirahabdulrahim@iium.edu.my)

This work was supported by the Asian Office of Aerospace Research and Development (AOARD) under Grant FA2386-23-1-4073 and Grant SPI23-179-0179.

ABSTRACT Satellite reliability is critical to ensuring uninterrupted operations in aerospace systems, where anomalies can lead to mission failures and significant economic losses. Existing anomaly classification methods often lack scalability, interpretability, and adaptability to diverse datasets. This study introduces the Trade-Space Exploration Machine Learning (TSE-ML) framework, a comprehensive pipeline for satellite anomaly classification that optimizes preprocessing, transformation, normalization, and machine learning stages. Leveraging a Seradata dataset spanning 66 years and 4,455 satellite records, the framework systematically evaluates four data cleaning methods, four data transformation techniques, five normalization strategies, and seven machine learning algorithms across 480 configurations. The optimal configuration, comprising Iterative Imputation, FastText, Robust Scaling, and Decision Tree, achieved the highest testing accuracy of 95.74% with competitive computational efficiency. The Decision Tree model delivered superior accuracy and provided interpretability, revealing critical factors influencing satellite anomalies, such as Age Since Launch, Design Life, and Orbit Category. Stratified 5-fold cross-validation ensured robustness and generalizability of the results. The TSE-ML framework's transparency and high performance enable actionable insights for improving satellite design, operational planning, and anomaly mitigation. Future research will focus on real-time anomaly detection, integrating satellite telemetry data, and extending the framework to other space applications. This study establishes a robust, interpretable foundation for advancing anomaly classification in aerospace engineering, addressing the dual challenges of reliability and operational efficiency.

INDEX TERMS Satellite anomaly detection, satellite reliability classification, trade-space exploration, data preprocessing techniques, machine learning models, seradata dataset, decision support systems.

I. INTRODUCTION

Satellite systems can be categorized based on their application areas, which include communications, earth observation and remote sensing, navigation, and research. They can also be classified according to their orbital paths: Low Earth Orbit (LEO), Medium Earth Orbit (MEO), Geosynchronous Earth Orbit (GEO), and High Elliptical Orbit (HEO), as illustrated in Fig. 1. Satellites vary widely in mass, from less than a kilogram to several tons, accommodating a diverse range of missions. Their operational lifespans depend on their mission objectives, with smaller satellites typically lasting a few years, while geostationary communication satellites may function for up to 15 years [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed M. A. Moustafa.

Suppose a critical component fails or an anomaly occurs during a satellite's operational lifetime. In that case, it may become partially or fully dysfunctional and remain in orbit unmanaged, contributing to the overcrowding of Earth's orbits [2]. This results in ''space junk'' accumulating in outer space. Currently, thousands of man-made objects are in orbit, with approximately 95% classified as ''space junk,'' represented by dots in Fig. 2. Each dot in the figure corresponds to an active satellite, an inactive satellite, or a piece of debris in LEO [3].

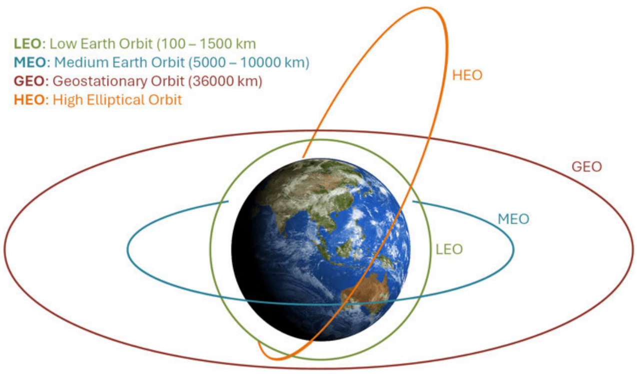

FIGURE 1. Various types of satellite orbits.

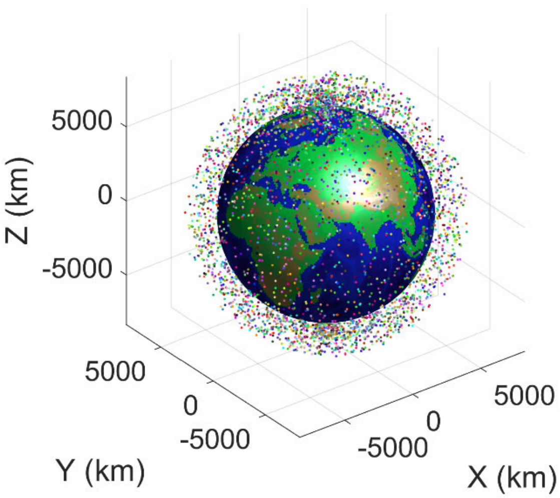

FIGURE 2. Simulated debris in LEO orbit.

As noted by Mital et al. [4], there are around 17,494 objects in Earth's orbital environment, including inactive satellites, debris, and rocket bodies. This accumulation has raised serious concerns, as satellites require stable conditions for effective anomaly management and resource optimization within satellite systems. Consequently, identifying and managing satellite anomalies, particularly regarding reliability, is critical to reducing satellite failures and promoting space sustainability. This issue highlights the need for responsible space exploration and the protection of space infrastructure.

In this research, we analyze Seradata, a leading database for satellite launches, reliability events, and space market analysis, to address the longstanding challenge of satellite anomalies. Over six decades, in-orbit anomalies have posed significant issues, as repairs are often impossible once satellites are deployed. Consequently, failed satellites become debris, floating in orbit and posing risks to active satellites and space assets [5]. Improving mission success rates requires a robust approach to satellite anomaly reliability identification, a critical component on which this study focuses. We aim to develop an optimized satellite anomaly reliability classification framework by leveraging a Trade-Space Exploration (TSE) framework integrated with machine learning (ML) techniques.

Previous studies have demonstrated the versatility of the Trade-Space Exploration (TSE) framework across various applications, including the formulation of system requirements and lifecycle cost analysis for satellite investment [6], Whole System Trades Analysis Tool (WSTAT) fot identifying the optimal system confguration concerning performance, cost, and risk using multiobjective optimization as implemented by the US Army Ground Vehicle Systems Center (GVSC) and TradeStudio developed by the US Army Engineer Research & Development Center (ERDC), an Engineering Resilient Systems (ERS) tool suite [7]. While these works highlight the effectiveness of TSE in space system design and decision-making, none have explored its integration with machine learning for satellite anomaly classification and reliability assessment, which forms the core novelty of this study.

Traditional anomaly detection in satellites often relies on manual assessments and rule-based systems, which are time-intensive and susceptible to human error. With the rise of big data and advancements in ML, there is a significant opportunity to enhance the accuracy and efficiency of anomaly identification in satellite systems. While recent studies have demonstrated ML's potential in areas like satellite bus type identification, stability prediction, and anomaly detection [4], there remains a gap in systematically integrating data preprocessing with ML and TSE to maximize model performance and decision-making accuracy.

ML, a branch of artificial intelligence (AI), enables models to learn from data and derive valuable insights, making it increasingly popular in the satellite industry over the past two decades [8]. However, developing effective ML models for anomaly detection requires rigorous data preprocessing, as raw satellite data is often incomplete, inconsistent, and noisy. Data preprocessing addresses these issues by cleaning, formatting, and structuring the data, ensuring it is prepared for model training and reducing computational overhead. Proper preprocessing is essential for achieving accurate and reliable ML models, particularly in complex domains like satellite anomaly detection, where data quality can directly impact model performance.

Established in 2013, Seradata has documented satellite reliability events across orbital categories, including

LEO, MEO, GEO, and HEO. The database covers the entire satellite lifecycle—from order and construction to launch, final positioning, and observed failures, categorized by severity, subsystem, or equipment involved [9], [10]. Our analysis leverages a dataset of 4,455 records, spanning 1957 to 2023, containing diverse satellite anomaly data.

This study aims to identify and evaluate the optimal combination of data preprocessing techniques and ML models for classifying satellite anomaly reliability. To achieve this, we apply a Trade-Space Exploration framework to experiment with various preprocessing methods and Machine Learning algorithms, seeking configurations that maximize accuracy and efficiency. The key contributions of this paper include:

1. *Novel TSE-ML Framework*: Developing a systematic TSE-ML framework that integrates data preprocessing and model selection to optimize satellite anomaly classification. This approach addresses gaps in traditional, rule-based anomaly identification methods by offering a data-driven, adaptive solution.
2. *Comprehensive Evaluation of Preprocessing Techniques*: Detailed analysis of multiple data preprocessing methods—including imputation, transformation, and normalization—demonstrating how specific preprocessing techniques impact model accuracy and processing time in satellite anomaly detection.
3. *Optimization of Accuracy and Efficiency*: Identifying optimal configurations that achieve high accuracy while minimizing processing time is crucial for real-time or near-real-time anomaly detection in satellite operations.
4. *Practical Framework for Reliability Assessment*: Provision of a scalable, data-driven framework for satellite operators and stakeholders, enabling proactive anomaly management and improving the reliability of satellite missions across diverse orbital and mission parameters.
5. *Empirical Validation with Large-Scale Data*: The proposed framework was validated on a comprehensive dataset from Seradata encompassing satellite anomalies over several decades, establishing its effectiveness and potential for broader application in the space industry.

These contributions underscore the TSE-ML framework's effectiveness in advancing satellite anomaly classification, providing a practical tool for enhancing satellite reliability and safety. The paper is structured as follows: Section II reviews existing data preprocessing and ML methods, space missions, and the TSE framework. Section III outlines the methodology, including data collection and ML models. Section IV covers the experimental setup and implementation details. Section V presents results and discussions, evaluating performance across TSE design scenarios. Section VI concludes with key findings, contributions, and suggestions for future work.

II. ADVANCEMENTS IN SATELLITE ANOMALY DETECTION, RELIABILITY, AND TRADE-SPACE EXPLORATION

Satellite anomaly detection is critical in ensuring operational reliability and longevity of space assets. Traditional approaches, often reliant on manual assessments and rule-based systems, have limitations in scalability and accuracy, particularly with the increasing complexity of satellite systems and data volumes. Recent advancements in machine learning and data-driven techniques offer new opportunities to automate and enhance anomaly detection. Yet, challenges remain in optimizing these methods for real-time, high-accuracy applications in satellite reliability.

A. OVERVIEW OF SATELLITE ANOMALIES AND RELIABILITY CHALLENGES

In the evolving satellite industry, operational failures that are not promptly and accurately managed can lead to severe malfunctions and significant financial losses [11]. Over the past six decades, records indicate that up to 200 satellites have experienced failures globally, highlighting the critical need for robust reliability strategies [12].

Satellite anomalies refer to unexpected deviations from normal satellite operations that can severely impact performance, efficiency, reliability, and lifespan [10]. These anomalies can result from various factors, including hardware and software malfunctions, operational errors, and environmental conditions like space debris, technical faults, or intentional interference and cyberattacks [10]. Understanding and managing these anomalies is essential for the success of satellite missions.

Notable incidents underscore the importance of anomaly management. In 2009, an Iridium communication satellite collided with a defunct Russian satellite, generating over 2,500 pieces of debris and contributing to the more than 18,000 artificial objects now monitored in Earth's orbit by the United States Space Surveillance Network [3]. In 2010, Intelsat's Galaxy-15 satellite, labeled as a "zombiesat," lost communication with ground control but continued to transmit signals, drifting in geosynchronous orbit and posing risks to nearby satellites for several months [5]. More recently, in 2021, an uncontrolled re-entry of China's Long March 5B rocket raised safety concerns among satellite operators [13], while in 2023, Russia's Luna-25 spacecraft crash landing on the Moon highlighted the potential risks of mission anomalies [14].

The growing complexity and volume of satellite data demand advanced, data-driven anomaly detection to ensure mission success. Traditional manual monitoring struggles to keep pace, increasing risks of collisions, service disruptions, and asset loss. Machine learning, combined with effective data preprocessing, refines raw data and detects subtle anomaly patterns with high accuracy. By integrating these technologies, satellite operations can shift toward real-time monitoring, proactive anomaly management, and improved decision-making. This scalable approach enhances satellite reliability, extends operational lifespans, and mitigates mission failures in increasingly congested orbital environments.

B. ADVANCED IN AI AND ML FOR SATELLITE OPERATIONS

As the aerospace industry evolves, spacecraft systems become increasingly complex, creating a growing demand for sophisticated anomaly detection solutions that effectively leverage satellite log data from comprehensive sources like Seradata. This log data, which contains detailed records of satellite anomalies, operational events, and failure instances, provides valuable insights into satellite behavior. By analyzing this data with AI-driven methods, satellite operators can automatically monitor and assess the operational status of satellites in orbit, allowing for rapid detection and mitigation of potential issues. These AI-enhanced methods offer significant practical value by providing real-time operational insights that support proactive anomaly management, thereby improving satellite reliability and mission success [12]. Additionally, there is a growing demand to minimize the overall cost of satellite operations, and automating maintenance and satellite management with AI is a promising approach to achieving substantial cost savings [8].

ML, a branch of artificial intelligence closely related to computational statistics, focuses primarily on making predictions and identifying patterns within large datasets [15]. Current ML research spans various fields, including natural language processing, computer vision, pattern recognition, cognitive computing, and knowledge representation, with predictive modeling being a core application in industrial settings [15]. ML techniques are categorized into several types: supervised, unsupervised, semi-supervised, and reinforcement learning. In supervised learning, models are trained on labeled data to predict new inputs accurately. Unsupervised learning, in contrast, is applied to unlabeled data to uncover hidden patterns or structures within the dataset. Semi-supervised learning combines labeled and unlabeled data to enhance learning accuracy, especially when labeled data is scarce. Reinforcement learning involves an agent interacting with its environment and learning to maximize rewards through feedback from its actions, making it ideal for dynamic, decision-driven scenarios.

Together, AI and ML advancements have significantly enhanced satellite anomaly classification capabilities. AI-driven techniques enable the efficient analysis of vast, complex datasets, while specialized ML models—ranging from supervised to reinforcement learning—allow for precise and flexible anomaly detection in diverse operational contexts. These advancements provide satellite operators with powerful tools for real-time monitoring, predictive maintenance, and proactive anomaly management, helping to prevent costly satellite malfunctions and extend operational lifespans. By leveraging the strengths of AI and ML, the aerospace industry is increasingly equipped to handle the growing demands of complex satellite systems, ensuring higher levels of reliability and mission success.

C. TRADE-SPACE EXPLORATION IN ENGINEERING DESIGN

Trade-Space Exploration (TSE) is a systematic approach used to evaluate a range of possible solution alternatives, encompassing a set of program parameters, system attributes, and performance characteristics. TSE is particularly valuable in complex engineering systems, where it balances trade-offs between cost, schedule, risk, and performance to meet specific standards and requirements [6]. This exploration space includes a wide spectrum of design choices, from overarching strategies to fine-tuned adjustments, allowing researchers to compare design scenarios and identify optimal solutions that satisfy often conflicting objectives.

As a core technique in engineering design, TSE provides a structured framework for navigating intricate design decisions where trade-offs are inevitable [7]. By examining multiple dimensions of viable solutions, TSE methodologies enable a thorough assessment of each option's strengths and limitations, facilitating an informed selection of the most effective configuration for a given system. TSE thus serves as a powerful tool for researchers to systematically evaluate and select design options, ensuring the chosen solution aligns with project goals and constraints [6].

TSE aims to optimize a solution by balancing multiple parameters or objectives, such as classification accuracy, processing time, and computational cost in satellite anomaly detection. Mathematically, TSE can be formulated as a *multi-objective optimization problem* where we seek an optimal set of parameters θ for the satellite anomaly classification model. The parameters may include preprocessing techniques, machine learning algorithms, and hyperparameters.

Let $f_1(\theta)$ represents classification accuracy and $f_2(\theta)$ represents processing. The TSE problem can be formulated as:

$$Optimize\ F(\theta) = [f_1(\theta), f_2(\theta)] \tag{1}$$

subject to constraints $C(\theta)$, such as operational costs and resource limits. The solution to this problem is a *Pareto optimal set* Θ^*, which includes configurations θ^*, where improving one objective, like accuracy, would degrade another, such as processing time.

$$\Theta^* = \left\{\theta \in \Theta \mid \nexists\, \theta' \in \Theta : \mathrm{F}\left(\theta'\right) \geq \mathrm{F}(\theta)\right\} \tag{2}$$

TSE is essential for optimizing configurations that balance accuracy, efficiency, and computational cost in satellite anomaly classification. TSE identifies optimal configurations to maximize detection accuracy while minimizing resource use by systematically exploring preprocessing techniques, algorithms, and model parameters. This approach is critical for real-time anomaly detection, preventing mission failures and costly damages. Leveraging TSE ensures strategies align with technical and operational demands, enhancing the robustness and reliability of satellite systems in congested orbital environments.

Compared to conventional model-based engineering (MBE) methods, Trade-Space Exploration (TSE) provides a more comprehensive, adaptable, and stakeholder-driven approach to system design and evaluation [6], [7]. While traditional MBE relies on predefined models and constrained design parameters, TSE allows for broader design trade-offs, enabling dynamic exploration of multiple configurations under varying constraints. Integrating machine learning (ML) further amplifies TSE's capabilities by enhancing computational efficiency, improving classification accuracy, and enabling data-driven decision-making [4], [16]. This synergy between TSE and ML results in a more robust and scalable framework, making TSE-ML particularly well-suited for applications such as satellite anomaly classification, where adaptability and predictive accuracy are critical.

D. SATELLITE RELIABILITY

Reliability is a critical design consideration for systems operating in hostile and inaccessible environments, such as satellites in orbit. Since satellite systems must function in extreme conditions without the possibility of in-situ maintenance, reliability measures are paramount in the design phase to mitigate the risk of functionality loss due to unexpected failures [1]. For high-value assets like satellites, where physical access for repair or adjustment is virtually impossible, the design must prioritize durability and resilience, ensuring the satellite can operate independently over its intended lifespan [17]. This makes reliability a foundational quality for satellite systems, as it directly impacts mission success and the long-term viability of space operations.

In engineering terms, reliability is the probability of a system performing its intended function under specified conditions for a particular period [17]. This translates to the probability of completing the assigned mission within a designated timeframe for satellite systems despite the hostile space environment. Reliability in satellite design is characterized by several factors, including the ability to perform consistently over time, withstand various stresses, balance reliability against other desirable qualities, achieve target performance within budget constraints, and maximize the satellite's utility once deployed [17]. The emphasis on reliability in satellite design is even more significant due to the high costs and complexities associated with satellite launch and maintenance, where a single anomaly or failure can jeopardize an entire mission and result in substantial financial losses [9].

Reliability, in the context of satellite anomaly classification, represents the probability that the system will function without failure over a given period or under specific conditions. In mathematical terms, reliability $R(t)$ is defined as:

$$R(t) = P(T > t) \tag{3}$$

where T is a random variable representing the time until a failure occurs, and t is the time under consideration. In anomaly detection, we can enhance reliability by optimizing the TSE parameters to maximize the detection of potential failures before they manifest in critical damage. By integrating anomaly detection models with high reliability, we aim to maximize $R(t)$ over the satellite's operational lifespan, ensuring it completes its mission without significant malfunctions.

E. SELECTION OF SERADATA FOR SATELLITE ANOMALY ANALYSIS

Seradata is an industry-leading open-source intelligence database that provides comprehensive information on international satellite and space launch activities, including detailed records of anomalies and failures [9]. Covering data from 1957 to 2023, Seradata documents around 4,455 instances of spacecraft anomalies, making it an invaluable resource for analyzing trends and patterns in satellite reliability and failure modes. The database categorizes failures into nine primary categories—attitude control, power, payload instrument, beam, control processor, telemetry, thermal, and transponder—providing a robust foundation for examining the diverse factors contributing to satellite malfunctions [14].

The selection of Seradata for this research is justified by its depth, historical range, and comprehensive anomaly reporting, which are essential for a rigorous analysis of satellite reliability. With over six decades of documented data, Seradata offers a broad view of satellite performance and failure characteristics, enabling researchers to understand the long-term impacts of design choices, operational conditions, and external environmental factors on satellite reliability. This extensive dataset allows for a granular exploration of failure modes, facilitating the development of ML models that can accurately predict and classify anomalies. By leveraging Seradata, this research can draw on a rich and diverse dataset to enhance satellite anomaly detection and provide insights that support the design of more resilient and reliable space systems.

III. DATA PREPROCESSING IN MACHINE LEARNING

Data preprocessing is a crucial step in machine learning to enhance raw data quality, consistency, and efficiency to extract meaningful insights [18], [19]. In ML, data preprocessing refers to the methods used to prepare raw data, ensuring it is accurate, consistent, and suitable for building and training models [19]. Preprocessing improves the quality of the training process by refining the data before analysis, helping to achieve reliable and interpretable results. Key preprocessing steps in this research include data cleaning, transformation, and normalization, each tailored to optimize the dataset for satellite anomaly classification.

Data preprocessing is essential to minimize noise, handle missing values, and enhance the relevance of features for the classification task [16]. Each preprocessing step includes several techniques: in data cleaning, methods such as Elimination, Mean Imputation, KNN Imputation, and Iterative Imputation address missing values to reduce data inconsistencies [20], [21], [22]. Data transformation techniques, including Label Encoding, Word2Vec, FastText, and Sentence Transformer, convert raw data into formats suitable for ML model consumption. Finally, normalization methods—such as Z-Score, Min-Max Scaling, Robust Scaling, Vector Normalization, and Power Transformation—adjust the scale of data to improve model stability and performance.

Exploring diverse data preprocessing techniques is essential for optimizing ML in satellite anomaly classification, where data quality directly influences model accuracy and reliability. Satellite data often contains noise, missing values, and inconsistencies due to challenging collection conditions, making robust preprocessing critical. Techniques such as imputation for missing values, encoding for categorical data, and normalization for feature scaling enhance data suitability for ML models, each impacting model performance differently. Identifying the optimal combination of these methods is crucial to achieving precise anomaly detection, reducing false positives, and ensuring reliable satellite operations. Proper preprocessing thus forms the backbone of an effective ML pipeline for robust, high-quality satellite anomaly classification.

A. DATA CLEANING FOR HANDLING MISSING VALUES

Data cleaning, or data cleansing, is the process of identifying, correcting, or removing errors and inconsistencies within a dataset to improve its quality and suitability for analysis [23]. One critical aspect of data cleaning is handling missing values, which can be managed through various methods such as elimination, mean imputation, k-Nearest Neighbors (KNN) imputation, and iterative imputation [21], [22].

1) ELIMINATION METHOD

The elimination method, also known as deletion, involves removing data entries (rows or columns) with missing values [21]. While this method is straightforward and effective for large datasets where missing data is not informative, it may result in significant information loss in smaller datasets [22]. In this study, listwise deletion is implemented, removing any row with missing values if the missing rate exceeds a set threshold T. This approach is formalized as in Eq. (4):

$$\textit{Eliminate Feature } X = \left\{ \textit{Remove } X \textit{ if } \frac{M_X}{N} > T \right. \tag{4}$$

where M_X represents the number of missing values in the feature X, N is the total number of observations, and T is the predefined threshold [22].

2) MEAN IMPUTATION

Mean Imputation replaces missing values in an attribute with the mean value of the observed data for that attribute. This method is computationally efficient, making it suitable for datasets with minimal missing data. However, it may not be ideal for datasets with large gaps, as it can introduce bias. Mean imputation is defined as:

$$\hat{X}_{ij} = \frac{1}{n_k} \sum\nolimits_{i:X_{ij} \in C_k} X_{ij} \tag{5}$$

where n_k is the number of non-missing values in the j-th feature of the class C_k [21].

3) KNN IMPUTATION

The KNN Imputation method fills in missing values based on the values of the k-nearest neighbors determined by distance measures such as Euclidean, Minkowski, Manhattan, and Cosine distances [24], [25]. Euclidean distance is commonly used due to its efficiency and effectiveness. KNN imputation with Euclidean distance is calculated as:

$$Dist_{xy} = \sqrt{\sum\nolimits_{k=1}^{m} \left(X_{ik} - X_{jk}\right)^2} \tag{6}$$

where $Dist_{xy}$ is the Euclidian distance between instances x and y, k denotes the number of attributes, X_{ik} is the value of the k-th attribute with missing data, and X_{jk} is the complete data value for the same attribute [25]. While effective, KNN imputation can be computationally intensive as it searches the entire dataset.

4) ITERATIVE IMPUTATION

Iterative Imputation models each feature with missing values as a function of other features in a round-robin fashion, using predictive modeling to estimate missing values [24], [26]. This method is particularly advantageous for multivariate datasets, leveraging correlations among features to provide more accurate imputations. The process of Iterative Imputation can be described as follows:

- Initialization: Start with an initial imputation for missing values, often using mean, median, or another simple method for each feature with missing values.
  - $X = \{X_1, X_2, \ldots, X_m\}$ represents the dataset with m features, some of which contain missing values.
  - X_i denotes the i-th features, and X_{-i} represents all other features, including X_i.
- Round-Robin Modeling:
  - For each feature X_i with missing values, construct a regression model f_i that predicts X_i based on the other features X_{-i}.
  - Let $\hat{X}_i^{(t+1)}$ denote the updated value of X_i at iteration $t+1$, predicted from the values of $X_{-i}^{(t)}$ from the previous iteration.
- Prediction: Use the model f_i to predict missing values in X_i based on the observed and currently imputed values of X_{-i}:

$$\hat{X}_i^{(t+1)} = f_i\left(X_{-i}^{(t)}\right) \tag{7}$$

where $\hat{X}_i^{(t+1)}$ is the imputed values for X_i based on the other features.

- Iteration and Convergence: Repeat the above process for each feature with missing values. Continue iterating until the imputed values stabilize or a predefined convergence criterion (such as a threshold for changes between successive imputations) is met.

After convergence, the dataset X has all missing values replaced with imputed values derived from the iterative process. The final imputed values for each missing entry in X are thus based on correlations with other features, leveraging multivariate relationships within the dataset.

Various data cleansing methods are essential for satellite anomaly classification due to satellite data's complex and often imperfect nature, which can contain significant noise, missing values, and inconsistencies from environmental interference, hardware limitations, and operational errors. Each data cleansing technique offers distinct strengths and weaknesses, making it critical to tailor approaches to specific data characteristics. For instance, elimination methods are straightforward and reduce noise by removing incomplete records, but they risk discarding valuable information, especially in smaller datasets. Mean imputation is computationally efficient and works well for datasets with minimal missing values but can introduce bias in more complex data patterns. KNN imputation, which infills missing values based on similar instances, preserves data structure more effectively but can be computationally intensive, especially for large datasets, due to its reliance on distance calculations. Iterative imputation, meanwhile, leverages multivariate relationships to predict missing values with high accuracy but requires significant computational power and can be sensitive to correlations among features. Given the unique challenges posed by satellite data, including the need for high accuracy in anomaly detection and the constraints of large, high-dimensional datasets, combining multiple cleansing techniques helps to achieve a balance between data completeness, computational efficiency, and accuracy, thereby optimizing the quality of data used in machine learning models for reliable satellite anomaly classification.

B. DATA TRANSFORMATION AND ENCODING METHODS

Data transformation, also known as encoding, converts non-numeric values into numeric representations, enabling ML algorithms to process categorical data effectively [27]. This involves altering data format, structure, or representation to make it more suitable for analysis and modeling. The choice of transformation technique varies based on the data type and the ML model's specific requirements. This research uses label encoding and various natural language processing (NLP) methods for data transformation, including Word2Vec, FastText, and Text Transformers.

1) LABEL ENCODING

Label Encoding is a straightforward technique provided by the Scikit-Learn library that transforms categorical variables into integer values ranging from 0 to $k_{classes} - 1$ [24], [27]. If X is the categorical variable with n samples, k represents the number of unique categories in X, which will be referred to as $k\_classes$ to keep it distinct and $Categories = \{C_1, C_2, \ldots, C_{k_{classes}}\}$ represents the unique categories within X, then label the encoding function $LE(C_i)$ can be defined as:

$$LE(C_i) = i - 1 \text{ for } i = 1, 2, \ldots, k_{classes} \tag{8}$$

This approach is particularly practical for machine learning models that interpret ordinal relationships, as it converts categorical data into a numeric format without introducing additional complexity. Label Encoding allows ML algorithms to process categorical data directly, enhancing model efficiency and simplifying the preprocessing pipeline by mapping each category to a distinct integer.

2) Word2Vec

Word2Vec is an NLP method representing words as vectors in a continuous vector space. It employs two primary model architectures: the Continuous Bag-of-Words (CBOW) and Skip-Gram (SG) models. CBOW predicts the current word based on its surrounding context, while Skip-Gram predicts surrounding words given a specific word [28], [29]. The objective function for training Word2Vec, commonly applied to optimize word embeddings, is defined as follows:

$$E = -\log\sigma\left(V'^{T}_{\omega_0}h\right) - \sum\nolimits_{\omega_{j\in W}} \log\sigma\left(V'^{T}_{\omega_j}h\right) \tag{9}$$

where ω_0 represents the target (output) word, V'_j is the output vector, and h denotes the hidden layer output, which is defined as $h = \frac{1}{C}\sum_{c=1}^{C} V_{\omega_c}$ for the CBOW model and $h = V_{\omega_I}$ for the SG model [28], [29]. Typically, the default specification for the hidden layer in Word2Vec is one layer with 10 neurons, although common implementations use between 100 and 1000 dimensions for the embedding space, depending on the dataset size and computational constraints [29].

3) FASTEXT

FastText is another NLP method developed by Facebook AI Research that efficiently learns word and sentence representations. Unlike Word2Vec, FastText considers the character n-grams, making it especially effective for morphologically rich languages and for handling out-of-vocabulary words [30], [31], [32]. Similar to Word2Vec, FastText can use either the SG or CBOW architecture but differs in its use of subword embeddings. This character-level focus allows FastText to generate more flexible and robust word representations. Studies show that FastText achieves comparable performance to Word2Vec while operating faster and producing smaller model sizes, which is advantageous for resource-limited environments.

In FastText, each word w is represented as a bag of character n-grams. The embedding for a word w is computed as the average of its n-gram embeddings:

$$Embedding\left(w\right)=\frac{1}{\left|g\left(w\right)\right|}\sum\nolimits_{g\in g(w)} v_g \tag{10}$$

where $g\left(w\right)$ represents the set of n-grams for the word w, v_g denotes the vector representation (embedding) for an n-gram g within $g\left(w\right)$, and $\left|g\left(w\right)\right|$ is the number of n-grams for the word w.

Eq. (10) captures both the structure of the word and its subword components, allowing FastText to generate robust word representations even for words that may not appear frequently in the training corpus. The final word embedding is thus a composition of its character-level n-grams, making FastText more effective than traditional word-level embeddings, especially in tasks involving rare or out-of-vocabulary words.

4) TEXT TRANSFORMER

Text Transformers, such as Bidirectional Encoder Representations from Transformers (BERT), were developed by Google AI and are designed to understand the context by analyzing sentences in both directions [33]. BERT training involves two stages: pre-training, where the model learns from large amounts of unlabeled data through specific language tasks, and fine-tuning, adapted to labeled data from specific tasks. This two-step process allows BERT to achieve high accuracy across various NLP applications by leveraging global and local contexts [31].

Text Transformers generate embeddings by processing text in both directions and capturing contextual relationships between words. Their core mechanism relies on self-attention to compute word representations based on a sentence's surrounding context. The self-attention mechanism can be formulated as follows:

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

where $Q\in\mathbb{R}^{n\times d_k}$, $K\in\mathbb{R}^{n\times d_k}$, and $V\in\mathbb{R}^{n\times d_v}$ are the matrices for queries, keys, and values, respectively. $QK^T\left(d_k\right)^{-0.5}$ computes the attention scores for all word pairs, capturing dependencies between words in the sequence. The softmax function normalizes these scores to form attention weights, which are then used to compute weighted sums of the values V for each word.

This self-attention mechanism is repeated across multiple layers and attention heads in a Transformer model, allowing the model to capture nuanced contextual information across different positions in the text. The final embedding for each word thus reflects its contextualized representation based on the entire input sequence.

Various data transformation techniques are essential for optimizing machine learning models in satellite anomaly classification, especially given satellite data's complex and diverse nature. Satellite datasets often include non-numeric information, such as operational statuses, failure descriptions, and system alerts written in sentences. To effectively process this textual data, it must be transformed into a numeric format suitable for machine learning algorithms. Techniques like label encoding provide a straightforward approach for transforming categorical text data into integer values, but they lack nuance for capturing complex relationships in textual information. More advanced methods, such as Word2Vec, FastText, and Transformers, offer powerful tools for converting sentences into meaningful vector representations. Word2Vec and FastText are particularly useful for capturing semantic relationships between words. FastText has the added advantage of handling out-of-vocabulary words through character n-grams, which is crucial in technical datasets where unique terminology may appear. Transformers like BERT go even further by leveraging contextual information and understanding relationships within and across sentences, making them well-suited for nuanced anomaly detection tasks in satellite data. Each transformation technique has its strengths and weaknesses, with simpler methods being computationally efficient but potentially less accurate and more advanced methods offering deep insights at the cost of increased computational resources. Given the high-stakes nature of satellite operations, where precise anomaly detection can prevent costly malfunctions, selecting appropriate transformation techniques is critical to optimize data quality and model performance.

C. DATA NORMALIZATION

Data normalization is the process of scaling or transforming data to ensure consistency in format and structure, making it suitable for machine learning algorithms [34], [35]. Several data normalization methods are applied in this research, including Min-Max Scaling, Z-Score Normalization, Robust Scaling, Vector Normalization, and Power Transformation. Each method has unique characteristics and serves different purposes when preparing the dataset for analysis.

1) MIN-MAX SCALING

Min-max scaling applies a linear transformation to scale data into a specific range, typically between 0 and 1. This approach aims to maintain the distribution of the data while rescaling values based on their minimum and maximum occurrences. Min-max scaling is especially useful for preserving zero entries in sparse data [24]. The formula for Min-Max Scaling is:

$$x'=\frac{x-\min\left(x\right)}{\max\left(x\right)-\min\left(x\right)} \tag{12}$$

where x' is the normalized value, x is the original value, and $\min\left(x\right)$ and $\max\left(x\right)$ are the minimum and maximum values of x in the dataset.

2) Z-SCORE NORMALIZATION

Z-Score Normalization, also known as *Standard Scaling*, standardizes data by centering it around the mean and scaling it to unit variance. This method is useful when the data distribution needs to be normalized for ML algorithms sensitive to scale. The Z-score of a sample x is calculated as follows:

$$z = \frac{x - \mu}{\sigma} \tag{13}$$

where μ is the mean of the data, and σ is the standard deviation of the data [24].

3) ROBUST SCALING

Robust Scaling mitigates the impact of outliers by using the median and interquartile range (IQR) rather than the mean and standard deviation. This approach scales data based on the IQR, defined as the range between the first quartile Q_1 and the third quartile Q_3 of the data. The formula for Robust Scaling is:

$$x' = \frac{x - Q_2}{Q_3 - Q_1} \tag{14}$$

where x' is the scaled value, x is the original value, and Q_2, Q_1, and Q_3 are the median, first quartile, and third quartile values, respectively [34].

4) VECTOR NORMALIZATION

Vector Normalization is a process that scales each data sample to have a unit norm, typically setting the magnitude (or length) of each vector to 1. This approach is particularly effective when dealing with data in vector form, as it preserves the direction of each data point while standardizing its scale, making it easier for machine learning algorithms to process and compare samples consistently [24].

Given a vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, the normalized vector $\mathbf{x}'$ is computed by dividing each component of $\mathbf{x}$ by the vector norm (magnitude). The formula for vector normalization is:

$$\mathbf{x}' = \frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{\mathbf{x}}{\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}} \tag{15}$$

where $\mathbf{x}'$ is the normalized vector with a unit norm, $\|\mathbf{x}\|$ denotes the Euclidean norm (magnitude) of $\mathbf{x}$. This normalization ensures that $\left\|\mathbf{x}'\right\| = 1$, allowing each data point to have a consistent scale while retaining its original direction.

5) POWER TRANSFORMATION

Power Transformation is a family of parametric transformations that aim to make data more Gaussian-like, which can improve the performance of ML models that assume normally distributed data. One of the most common Power Transformations is the Box-Cox transformation, which was introduced by Box and Cox in 1964. The Box-Cox transformation formula is:

$$\psi^{BC}(\lambda, x) = \begin{cases} \dfrac{(x^{\lambda} - 1)}{\lambda} & \mathit{if}\ \lambda \neq 0 \\ \log(x) & \mathit{if}\ \lambda = 0 \end{cases} \tag{16}$$

where λ is a parameter that determines the nature of the transformation and x is the data value [36], [37]. Power transformations are beneficial for stabilizing variance and making data more symmetrical, thus aiding in model accuracy.

Data normalization is a crucial step in satellite anomaly classification, as it ensures that each feature contributes proportionally to the model's performance by scaling data to a typical range or distribution. Since satellite data often consists of measurements across various units and scales, such as temperature, voltage, and position, normalization prevents features with larger numerical ranges from dominating the learning process. Techniques like Min-Max Scaling and Z-Score Normalization standardize the range and distribution of the data, making it easier for machine learning models to converge and identify meaningful patterns in anomaly detection. Normalization enhances model accuracy and stability by reducing biases caused by differing feature scales, which is essential in high-stakes applications like satellite anomaly classification, where precision is critical. However, normalization can also be sensitive to outliers. At the same time, Min-Max Scaling may overemphasize extreme values. Techniques like Robust Scaling are more resilient, adjusting data based on interquartile ranges to mitigate outlier influence. Despite this, normalization has inherent trade-offs, as the choice of method depends on data characteristics and model requirements. When applied thoughtfully, normalization strengthens anomaly classification models, enabling them to reliably detect deviations in satellite behavior and support real-time monitoring, where quick, accurate anomaly identification is paramount for maintaining operational reliability in space systems.

IV. TSE-ML FRAMEWORK AND IMPLEMENTATION

This section details the framework's conceptual structure, key algorithmic steps, and technical setup, including hardware and software tools. Additionally, it provides an exploratory analysis of the Seradata dataset, establishing a data foundation for optimizing satellite anomaly classification through the TSE approach.

A. FRAMEWORK DESIGN

The proposed TSE (Trade-Space Exploration) framework aims to optimize satellite anomaly classification by balancing accuracy and processing time through a three-step approach: data preprocessing, machine learning model development, and performance evaluation, as illustrated in Figure 3. This workflow is further detailed in Figure 4, which presents the TSE algorithm, providing a step-by-step breakdown of each stage. Figures 3 and 4 offer a comprehensive view of the TSE framework's structure and operational flow, guiding the selection of preprocessing methods, machine learning models, and evaluation criteria to achieve optimal classification results.

The Data Preprocessing stage refines raw data from Seradata, addressing missing values and inconsistent scales to enhance model performance. This stage involves three key operations: data cleansing, transformation, and

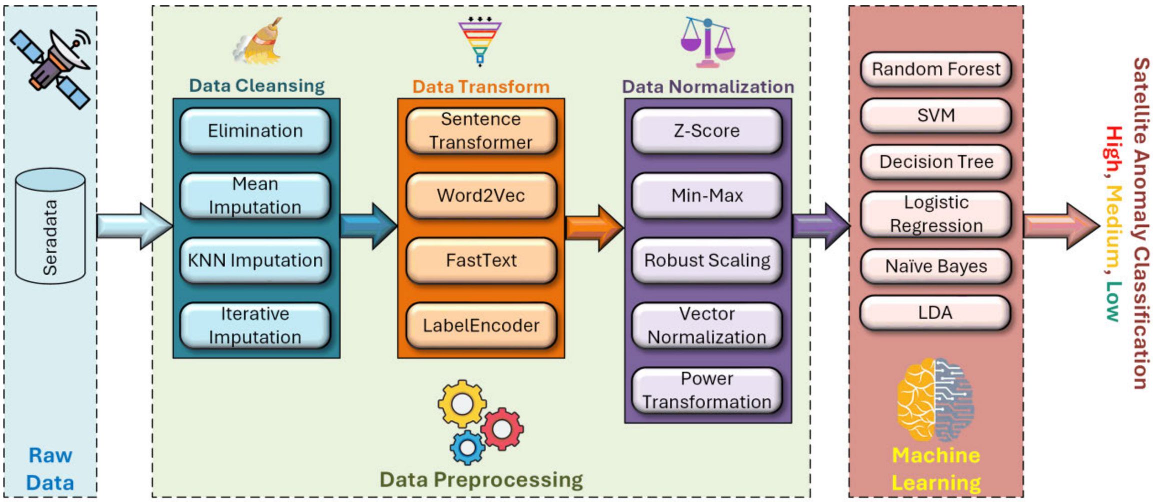

FIGURE 3. Proposed TSE framework.

Algorithm 1 TSE Framework for Satellite Anomaly Classification

```
Require: Raw satellite data D from Seradata with mixed
    data types and potential missing values
Ensure: Classified satellite anomaly level (High,
    Medium, Low) based on reliability calculation
Raw Data Collection
Collect raw dataset D from Seradata, containing various satellite attributes and anomaly records.
Data Preprocessing
Data Cleansing
for each feature f in dataset D do
  if missing values in f then
    Choose an imputation method:
    if Elimination then
      Remove rows with missing values (if feasible based on data size).
    else if Mean Imputation then
      Replace missing values in f with the mean of observed values.
    else if KNN Imputation then
      Replace missing values by averaging the k-nearest neighbors.
    else if Iterative Imputation then
      Predict missing values iteratively based on other features using a regression model.
    end if
  end if
end for
Data Transformation
for each categorical feature c in D do
  Transform non-numeric data using one of the following:
  Label Encoding, Word2Vec, FastText, or Sentence Transformers (e.g., BERT)
end for
Data Normalization
for each numeric feature n in D do
  Normalize feature values using one of the following methods:
  Min-Max Scaling, Z-Score Normalization, Robust Scaling, Vector Normalization, or Power Transformation
end for
Resulting processed dataset D' is cleaned, transformed, and normalized.
Machine Learning
Split D' into training and testing sets (e.g., 70:30 split).
Model Selection:
Define a set of candidate models: Random Forest, SVM, Decision Tree, Logistic Regression, Naive Bayes, LDA.
Model Training and Optimization:
for each model M in the candidate set do
  Train M on the training set.
  Perform cross-validation to tune hyperparameters, optimizing for classification accuracy and processing time.
end for
Select the model M* with the best trade-off between accuracy and processing time.
Satellite Anomaly Classification
Evaluate M* on the testing set.
Predict anomaly levels as High, Medium, or Low based on reliability calculation.
Optimization Objective
The TSE framework is designed to maximize classification accuracy and minimize processing time for efficient and reliable anomaly detection.
return Anomaly classification for each satellite (High, Medium, Low), optimized for accuracy and processing efficiency.
```

FIGURE 4. Proposed TSE algorithm.

normalization. Data cleansing employs methods such as mean imputation for simplicity, KNN imputation for accuracy across mixed datasets, and iterative imputation for robust handling of complex missing patterns. In data transformation, techniques like Label Encoding, Word2Vec, FastText, and Sentence Transformers convert categorical or textual data into numerical formats that ML models can process efficiently. Finally, data normalization methods such as Min-Max scaling, Z-Score normalization, and Vector Normalization ensure that features are scaled appropriately, addressing the sensitivity of AI models to varied data ranges.

The Machine Learning (ML) Model Development step selects and trains models for anomaly classification. Key algorithms evaluated include Random Forest, Support Vector Machines (SVM), Decision Tree, K-nearest neighbors (k-NN), Logistic Regression, Naive Bayes, and linear discriminant analysis (LDA). Models are trained on a 70:30 train-test split, which provided the best balance between performance and reliability across our experiments.

The Performance Evaluation and Optimization phase identifies the optimal TSE configuration by systematically comparing model performance metrics. This approach uses accuracy and processing time as the core optimization objectives, ensuring each model configuration meets the operational demands for real-time anomaly detection in satellite operations. The selected configuration is then evaluated against predefined reliability thresholds to classify anomalies as high, medium, or low impact, aiding proactive decision-making in satellite mission management.

Through this framework, the TSE approach enables an in-depth exploration of various model and preprocessing configurations, balancing the trade-offs between processing resources and detection accuracy. This multi-dimensional evaluation ensures the TSE framework is optimized to handle the operational complexities of satellite anomaly detection, providing a robust, scalable solution for enhanced reliability in increasingly congested orbital environments.

B. IMPLEMENTATION SETUP

The experimental setup for this research is designed to ensure the replicability and robustness of the TSE framework in satellite anomaly classification. The experiments were conducted on a dedicated, high-performance server with minimal background applications to eliminate external variables and ensure consistent results. The choice of hardware and software reflects the demanding computational requirements of large-scale satellite data analysis and machine learning model optimization.

1) HARDWARE CONFIGURATION

Table 1 details that the hardware configuration includes state-of-the-art components optimized for intensive computational tasks. The Intel Core i9-14900K processor, operating at a clock speed of 3.20 GHz, ensures high processing power and efficiency for data preprocessing and model training. The NVIDIA GeForce RTX 4090 GPU accelerates parallel processing, essential for complex ML algorithms and deep learning models. With 64 GB of DDR5 RAM, the system supports large-scale memory-intensive operations, including handling the extensive Seradata dataset. Storage is divided between a 500 GB SSD for high-speed access to frequently used files and a 6 TB HDD for long-term storage, enabling seamless data management.

TABLE 1. Hardware specifications.

Category	Specification
CPU	Intel(R) Core i9-14900K, 3.20 GHz
GPU	NVIDIA GeForce RTX 4090
RAM	64.0 GB DDR5
Storage (SSD)	500 GB
Storage (HDD)	6 TB

2) SOFTWARE AND LIBRARY SETUP

The software environment, outlined in Table 2, was curated to align with the TSE framework's computational needs, ensuring speed, efficiency, and reproducibility. Python 3.11.9 is the foundation, providing extensive libraries and tools for data analysis and machine learning. GPU acceleration, enabled by CUDA 12.4 and PyTorch 2.4.0, significantly reduces training time for ML models. Essential libraries like pandas and Numpy handle large-scale data preprocessing tasks, while Scikit-learn enables the implementation of diverse ML algorithms.

Advanced text processing tools, including gensim, fasttext, and sentence-transformers, transform unstructured textual data into numeric representations, optimizing it for ML model consumption. Seaborn and Matplotlib provide intuitive plotting capabilities for data visualization, aiding in exploratory data analysis and performance evaluation.

TABLE 2. Software and library specifications.

Category	Version
Windows Server Datacenter	2022
WSL Ubuntu	24.04 LTS
Jupyter-Lab	4.2.5
Python	3.11.9
pandas	2.2.2
numpy	1.26.4
sklearn	1.5.1
seaborn	0.13.2
matplotlib	3.8.4
genism (word2vec)	4.3.3
fasttext	0.9.3
sentence-transformers	3.0.1

3) MACHINE LEARNING CONFIGURATION

As detailed in Table 3, the machine learning models were configured with carefully chosen hyperparameters to ensure fair and robust evaluations. Default configurations maintained consistency and focused on preprocessing techniques and trade-space exploration. These settings were sufficient to achieve high accuracy and efficiency, leveraging the optimized hardware and software setup.

Nevertheless, modifications to the default parameters were considered to accommodate scenarios where performance

TABLE 3. Hyperparameters setup.

ML	Parameters
Random Forest	RandomForestClassifier(n_estimators=200, random_state=0)
SVM	SVC(kernel='linear', random_state=7)
Decision Tree	DecisionTreeClassifier(criterion='gini', splitter: 'best', max_depth=None, min_samples_split=2, random_state=0)
K-NN	KNeighborsClassifier(n_neighbors=5, weights= uniform, algorithm='auto', leaf_size: 30, p=2, metric: 'minkowski', metric_params: Non, n_jobs: None)
Logistic Regression	LogisticRegression(penalty: 'l2', dual: False, intercept_scaling: 1, solver: 'lbfgs', max_iter: 100, random_state=0)
Naïve Bayes	GaussianNB(var_smoothing=1e-9)
Linear Discriminant Analysis	LinearDiscriminantAnalysis(solver: 'svd', tol: 0.0001)

metrics such as accuracy or processing time fell below acceptable thresholds. These adjustments were made iteratively, based on performance diagnostics, to enhance model outcomes while preserving computational efficiency. Such parameter tuning ensured that the models were adaptable to diverse data characteristics, optimizing their ability to classify satellite anomalies accurately and reliably. This adaptive approach allows the framework to remain flexible while addressing potential underperformance, ensuring the robustness and scalability of the machine learning pipeline across varying conditions.

The choice of hyperparameters enhances the balance between computational feasibility and model robustness, facilitating reliable and scalable satellite anomaly classification. For performance evaluation, this setup prioritizes accuracy over training and testing time. This prioritization is intentional, as future advancements in hardware technology can readily address processing speed. However, achieving high classification accuracy is critical to the framework's success and applicability in real-world satellite reliability assessments. By ensuring that the models are tuned for maximum precision, the study sets a strong foundation for operational reliability, even as technology evolves.

C. DATASET PREPARATION AND EXPLORATORY DATA ANALYSIS

The data for this study was sourced from Seradata, a proprietary database known for its comprehensive records of satellite operations and anomalies. Covering 66 years, from 1957 to 2023, the dataset comprises 4,455 entries, offering a detailed view of satellite anomaly events and associated parameters. The features within the dataset, listed in Table 4, include both numerical and non-numerical attributes, such as the satellite's design life, operational orbit, and mission type, providing a diverse set of inputs for machine learning models.

From this dataset, 3,050 records were selected for further analysis after excluding retired satellites. This selection criterion ensures a focus on unreliable entries, which is critical for identifying patterns and insights into satellite anomalies. By concentrating on data with known reliability concerns, this study seeks to uncover relationships between key attributes, such as bus type, mission, orbit, and satellite mass, and the occurrence of anomalies.

TABLE 4. Seradata satellite details.

Feature	Numeric	Description
Event (t)	No	Failure reason category
Age Since Launch At Event (Years) (t)	Yes	Number of years elapsed from the satellite's launch to the failure event
Design Life (years)	Yes	Expected operational lifespan of the satellite
Bus Type	No	The structural and functional core of the satellite, housing payload and instruments
Orbit Category	No	The satellite's orbital classification, including LEO, MEO, GEO, and HEO
Mission - Primary	No	The primary mission or objective of the satellite
Mass At Launch (kg)	Yes	The total mass of the satellite, including payload and instruments

1) DATA TRANSFORMATION

Non-numerical features were transformed into numerical representations to facilitate machine learning model training. Table 5 provides an example of this process for a specific event, "Retired - Due to Unknown Anomaly, I," where text data was encoded into vectors using FastText, Word2Vec, and Sentence Transformers. These vectors were then averaged to produce a single numeric value, a widely used approach in clustering and semantic analysis tasks [38]. Averaging is computationally efficient and straightforward. It allows machine learning models to handle variable-length text inputs effectively by creating fixed-size representations.

If the resulting accuracy proves insufficient, adjustments will be made to the transformation process. Instead of averaging, alternative methods may be explored, including retaining the entire vector representation to preserve the detailed contextual information encoded by FastText, Word2Vec, and Sentence Transformers. While this approach will increase the computational requirements due to the larger feature size, it may enhance model performance by providing richer, more nuanced data for anomaly classification. This flexibility in transformation strategies ensures that the data preparation process can adapt to the needs of the machine learning model, prioritizing accuracy and robustness in the final results.

TABLE 5. Sample of encoding processes for non-numeric features.

Methods	Vector (100x100)	Numeric
FastText	[-0.0033399574, 0.019801503, -0.036131494, 0.0...	-0.000437
Word2Vec	[-0.03911728, 0.042212844, 0.08253286, 0.08839...	0.012623
Sentence Transformer	[-0.042186722, 0.008450047, 0.013887661, 0.049...	-0.00071

By transforming textual features into fixed-size numeric representations, the dataset becomes suitable for machine learning models, which typically require consistent input formats. This process ensures efficient computation while maintaining the semantic integrity of the original data.

2) RELIABILITY CALCULATION

The reliability of satellite anomalies is calculated using the formulation in Eq. (17), where NR represents the normalized reliability:

$$NR = \left| \frac{a - d}{d} \right| \tag{17}$$

where a the satellite's age from launch to failure, and d signifies the satellite's design life. A higher NR value indicates greater unreliability. This formulation is adapted from [17], who define reliability as a function of multiple characteristics, including durability, resistance to stress, and trade-offs between cost and performance. Reliability is categorized into three levels: low ($0 \leq NR \leq 40$), medium ($41 \leq NR \leq 70$), and high ($71 \leq NR \leq 100$).

3) CORRELATION ANALYSIS

A correlation matrix was constructed to analyze relationships between features, as shown in Figure 5. This matrix measures the strength and direction of linear relationships between variables, with values ranging from −1 to 1. Key observations include:

- **Age Since Launch and Event (t)**: Weak negative correlation (-0.08), suggesting events are slightly more likely earlier in a satellite's operational lifespan.
- **Design Life and Age Since Launch (t)**: Moderate positive correlation (0.5), indicating satellites with longer design lives tend to experience failures later in their lifespan.
- **Design Life and Event (t)**: Weak negative correlation (−0.15), suggesting satellites with longer design lifespans are marginally less prone to anomalies.
- **Bus Type and Orbit Category**: Weak correlation (0.21), indicating minor associations between satellite structure and orbital position.
- **Mission and Orbit Category**: Moderate positive correlation (0.41), implying the primary mission is somewhat influenced by orbital placement.
- **Mass at Launch and Event (t)**: Weak negative correlation (-0.13), suggesting heavier satellites may have a slightly lower likelihood of encountering anomalies.

This analysis reveals patterns and relationships among satellite attributes, aiding the development of machine-learning models tailored for anomaly classification.

V. RESULT AND DISCUSSION

This section presents the experimental results and critical Trade-Space Exploration (TSE) framework analysis. The experiments were conducted systematically to evaluate the impact of various preprocessing and machine learning methods on satellite anomaly classification accuracy and processing time. Each experimental phase focused on a specific pipeline component, building upon prior research [39] and incorporating improvements for more reliable performance. In the previous research, the best configuration for the TSE framework included elimination for data cleaning, Label Encoder for data transformation (as only Label Encoder was used in the earlier study), Z-Score for normalization, and Support Vector Machines (SVM) for the machine learning model. While this configuration achieved reasonable performance, its reliance on limited preprocessing and machine learning techniques revealed the need for a broader exploration of alternative methods. This study expands the evaluation by incorporating advanced techniques across the pipeline to improve accuracy and efficiency. Furthermore, the training and testing accuracies reported in this study were obtained using stratified 5-fold cross-validation, which ensures that each fold preserves the distribution of satellite anomalies, reducing bias and improving the generalizability of the results. This comprehensive approach seeks to identify a more robust configuration that outperforms the baseline framework.

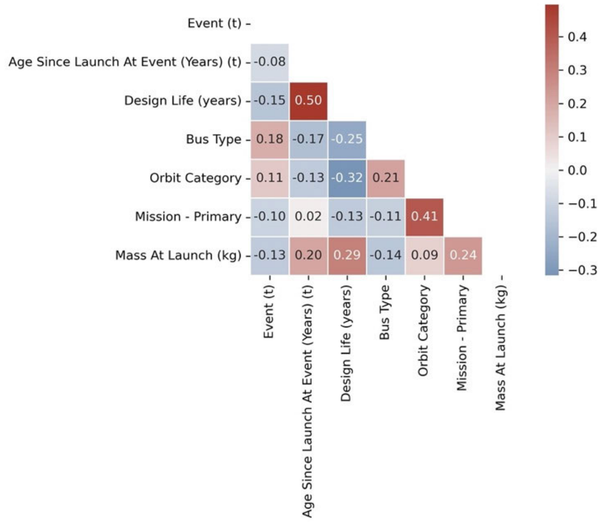

FIGURE 5. Correlation matrix heatmap.

A. EVALUATION OF DATA CLEANING METHODS

The first experiment in the TSE framework aimed to determine the most effective data-cleaning method for satellite anomaly classification. The experiment systematically evaluated four methods: Elimination, Mean Imputation, KNN Imputation, and Iterative Imputation, as depicted in Figure 6. These methods were assessed based on their training and testing accuracy as well as processing times, with results summarized in Table 6. This step was critical as data cleaning provides the foundation for the TSE pipeline's subsequent transformation and classification processes.

FIGURE 6. Data cleaning experiments.

TABLE 6. Data cleaning performance.

Data Cleaning	Training		Testing	
	Accuracy	Time (s)	Accuracy	Time (s)
Elimination	0.94643	0.09390	0.90678	0.12261
Mean Imputation	0.94754	0.08354	0.91694	**0.05131**
KNN Imputation	0.94754	0.08352	0.91694	0.08949
Iterative Imputation	0.94754	0.12447	**0.91803**	0.12929

Elimination, while computationally efficient with the shortest training (0.0939 seconds) and relatively fast testing times (0.1226 seconds), achieved the lowest testing accuracy (90.68%) due to the loss of critical information from removing incomplete records, making it less suitable for complex datasets like satellite anomaly classification. Mean Imputation and KNN Imputation both achieved identical testing (91.69%) and training (94.75%) accuracies, with Mean Imputation standing out for its exceptional computational efficiency, particularly with the shortest testing time (0.0513 seconds), making it ideal for resource-constrained scenarios. KNN Imputation, while slightly slower, maintained contextual integrity by considering relationships among neighboring data points, though it failed to outperform Iterative Imputation. Iterative Imputation emerged as the best method, achieving the highest testing accuracy (91.80%) by leveraging feature interdependencies through iterative refinement. Despite having the longest training (0.1244 seconds) and testing times (0.1293 seconds), its superior accuracy makes it the most reliable choice for high-precision tasks like satellite anomaly classification. Based on these results, Iterative Imputation will be selected as the optimal data-cleaning method for use in the optimal TSE experiment.

B. EVALUATION OF DATA TRANSFORMATION METHODS

The second experiment, illustrated in Figure 7, aimed to evaluate the impact of different data transformation methods on satellite anomaly classification performance. The methods tested included Label Encoder, Word2Vec, FastText, and Sentence Transformer. These methods were compared based on training and testing accuracy and processing time to identify the most effective approach for transforming non-numerical data into numerical representations suitable for machine learning models. The results are presented in Table 7.

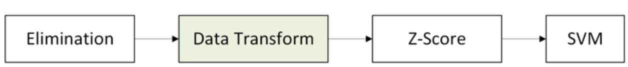

FIGURE 7. Data transform experiments.

Label Encoder, the simplest method, achieved the fastest training (0.0939 seconds) and testing times (0.1226 seconds) but had the lowest testing accuracy (90.68%) due to its inability to capture semantic relationships within the data, making it unsuitable for tasks requiring deeper understanding. Word2Vec offered a balance between accuracy and processing time, achieving a testing accuracy of 91.00% and demonstrating its ability to capture meaningful relationships, though its processing times (0.5079 seconds for training and 0.4869 seconds for testing) were higher. FastText, despite its significant computational cost (4.6013 seconds for training and 4.6403 seconds for testing), achieved the highest testing accuracy (91.12%), making it the most effective method for generalization and improving anomaly classification. Conversely, Sentence Transformer had the lowest testing accuracy (89.68%) and the highest computational cost, making it unsuitable for this context. Ultimately, FastText was selected as the optimal data transformation method for its superior accuracy, ensuring a strong data representation for use in the optimal TSE experiment.

TABLE 7. Data transform performance.

Data Transform	Training		Testing	
	Accuracy	Time (s)	Accuracy	Time (s)
Label Encoder	0.94643	0.0939	0.90678	**0.12261**
Word2Vec	0.94784	0.50796	0.91009	0.48694
FastText	0.94314	4.60136	**0.91116**	4.64031
Sentence Transf.	0.94503	48.5901	0.89688	48.5589

C. EVALUATION OF DATA NORMALIZATION METHODS

The third experiment evaluated the performance of various data normalization methods within the TSE pipeline, as depicted in Figure 8. The methods assessed included Z-Score, Min-Max Scaling, Robust Scaling, Vector Normalization, and Power Transformation. These methods were evaluated based on training and testing accuracy and processing times, with the results summarized in Table 8.

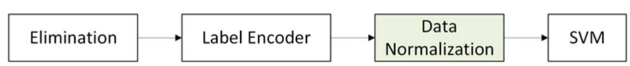

FIGURE 8. Data normalization experiments.

TABLE 8. Data normalization performance.

Data Transform	Training		Testing	
	Accuracy	Time (s)	Accuracy	Time (s)
Z-Score	0.94643	0.09390	0.90678	0.12261
Min-Max	0.76456	0.07513	0.72805	0.09964
Robust Scaling	0.94784	0.10210	**0.91116**	**0.06891**
Vector Norm.	0.42152	0.26186	0.38595	0.36671
Power Transform.	0.94502	0.07572	0.90459	0.09142

Robust Scaling emerged as the most effective normalization method, achieving the highest testing accuracy (91.12%) and the shortest testing time (0.0689 seconds), making it both accurate and computationally efficient. By leveraging the median and interquartile range, Robust Scaling effectively mitigates the influence of outliers, which is critical for satellite anomaly classification. In contrast, Z-Score normalization provided reasonable results with a testing accuracy of 90.68% and moderate processing times but was not the optimal choice. Min-Max Scaling, while the fastest method (0.0996 seconds for testing), had the lowest testing accuracy (72.81%) due to its sensitivity to outliers, making it suitable only for applications prioritizing speed over precision. Vector Normalization and Power Transformation underperformed in accuracy and efficiency, with Vector Normalization showing the poorest results overall. Therefore, Robust Scaling was selected as the optimal method for this study, ensuring the dataset is well-prepared for subsequent steps in the TSE pipeline and enhancing the reliability and accuracy of satellite anomaly classification.

D. EVALUATION OF MACHINE LEARNING METHODS

The fourth experiment assessed the performance of various machine learning (ML) algorithms in the TSE pipeline, as depicted in Figure 9. The algorithms evaluated included SVM, Random Forest, Decision Tree, KNN, Logistic Regression, Naïve Bayes, and Linear Discriminant Analysis (LDA). These methods were compared based on their training and testing accuracy and processing time to determine the most effective ML model for satellite anomaly classification. The results are summarized in Table 9.

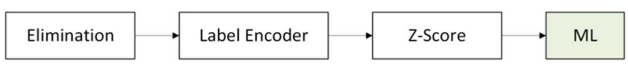

FIGURE 9. Machine learning experiments.

Decision Tree emerged as the optimal machine learning model for satellite anomaly classification, achieving the highest testing accuracy (94.96%) and the shortest testing time (0.0455 seconds), making it both effective and efficient. Its strong generalization ability and interpretable structure make it ideal for applications requiring high accuracy and rapid predictions. In comparison, SVM, the baseline model, delivered a reasonable performance with a testing accuracy of 90.68% and competitive processing times but lacked the generalization capabilities of superior models. Random Forest improved testing accuracy to 91.78% with moderate processing costs, while Logistic Regression provided a balance of accuracy (91.66%) and speed, making it a viable option for resource-constrained environments. KNN performed poorly, with a testing accuracy of 70.50%, due to its sensitivity to noisy data and higher processing time. In contrast, Naïve Bayes and LDA, with testing accuracies of 72.69% and 69.96%, respectively, proved unsuitable for the dataset's complexity. Based on its superior performance, the Decision Tree was selected as the primary ML method for subsequent experiments in the TSE pipeline.

TABLE 9. Machine learning performance.

Machine Learning	Training		Testing	
	Accuracy	Time (s)	Accuracy	Time (s)
SVM	0.94643	0.09390	0.90678	0.12261
Random Forest	0.94596	0.49721	0.91775	0.11001
Decision Tree	0.96241	0.04203	**0.94959**	**0.04552**
KNN	0.78572	0.04029	0.70501	0.19247
Logistic Regress.	0.94831	0.04540	0.91663	0.04584
Naïve Bayes	0.73307	0.04126	0.72692	0.04620
LDA	0.73259	0.05150	0.69956	0.04565

E. EVALUATION OF OPTIMAL TSE PIPELINE WITH VARIOUS MACHINE LEARNING MODELS

The next experiment evaluated the performance of different machine learning models within the optimized TSE pipeline, which consisted of Iterative Imputation, FastText for data transformation, and Robust Scaling for normalization, as illustrated in Figure 10. The models tested included SVM, Random Forest, Decision Tree, KNN, Logistic Regression, Naïve Bayes, and LDA, with results presented in Table 10.

FIGURE 10. Optimal TSE experiments.

TABLE 10. Optimal TSE experiments on various machine learning.

Machine Learning	Training		Testing	
	Accuracy	Time (s)	Accuracy	Time (s)
SVM	0.94333	2.24174	0.91475	2.23804
Random Forest	0.95504	2.44868	0.91803	2.25003
Decision Tree	0.97237	2.21030	**0.95738**	2.21143
KNN	0.73724	2.20933	0.71913	2.26901
Logistic Regress.	0.93864	2.21925	0.91913	2.21133
Naïve Bayes	0.74239	2.21046	0.70055	2.21141
LDA	0.70679	2.21291	0.70164	**2.21116**

Decision Tree emerged as the best-performing model in the optimized TSE pipeline, achieving the highest testing accuracy (95.74%) with competitive processing times (2.2114 seconds). Its robustness, interpretability, and superior generalization make it the most suitable choice for satellite anomaly classification. In comparison, SVM and Random Forest achieved respectable testing accuracies of 91.47% and 91.80%, respectively, but at higher computational costs, while Logistic Regression offered a balance of accuracy (91.91%) and efficiency, making it a viable alternative in resource-constrained scenarios. KNN underperformed with a testing accuracy of 71.91% and longer processing times, demonstrating sensitivity to noise and inefficiency in high-dimensional datasets. Naïve Bayes and LDA, with the lowest testing accuracies (70.05% and 70.16%, respectively), proved unsuitable for the dataset's complexity. Overall, the Decision Tree's combination of accuracy and computational efficiency reaffirms it as the most reliable model for the TSE pipeline.

TABLE 11. Top 10 Optimal method combinations for the TSE framework, Including data cleaning, Transformation, Normalization, and Machine learning, with corresponding training and testing performance metrics.

Data Cleaning	Data Transform	Data Normalization	Machine Learning	Training		Testing	
				Accuracy	Time (s)	Accuracy	Time (s)
KNN Imputation	Label Encoder	MinMax	Decision Tree	0.97237	0.06928	**0.96066**	0.07460
KNN Imputation	Label Encoder	Z-Score	Decision Tree	**0.97564**	0.06223	0.95956	0.06327
KNN Imputation	FastText	Z-Score	Decision Tree	0.97330	3.78462	0.95956	3.78481
KNN Imputation	Word2Vec	Robust Scaling	Decision Tree	0.96956	0.52795	0.95847	0.53160
Iterative Imputation	Label Encoder	Robust Scaling	Decision Tree	0.97283	0.07298	0.95847	0.07244
KNN Imputation	Label Encoder	Robust Scaling	Decision Tree	0.97002	0.10600	0.95738	0.10517
Iterative Imputation	FastText	Robust Scaling	Decision Tree	0.97237	2.21030	0.95738	2.21143
Mean Imputation	Word2Vec	MinMax	Decision Tree	0.96909	0.54069	0.95628	0.54091
Mean Imputation	Word2Vec	Z-Score	Decision Tree	0.97237	0.64938	0.95628	0.65085
Mean Imputation	Label Encoder	Robust Scaling	Decision Tree	0.97330	**0.02685**	0.95628	**0.03265**

F. OPTIMAL METHOD COMBINATIONS FOR TSE

The final experiment sought to identify the most effective combination of preprocessing methods and machine learning models for the Trade-Space Exploration (TSE) framework by testing all possible configurations. This evaluation integrated various techniques across four critical stages: data cleaning, data transformation, data normalization, and machine learning. The goal of systematically exploring these combinations was to determine the configuration that maximizes accuracy and computational efficiency for satellite anomaly classification. As shown in Fig. 3, with four methods of data cleaning, four methods of data transformation, five methods of data normalization, and seven machine learning models, a total of 480 possible combinations were evaluated. The results of the top 10 combinations are summarized in Table 11, showcasing the configurations that achieved superior performance in training and testing accuracy while maintaining competitive processing times.

The results reveal that the combination of KNN Imputation for data cleaning, Label Encoder for data transformation, MinMax normalization, and Decision Tree for machine learning achieved the highest testing accuracy (96.07%) while maintaining an efficient testing time of 0.0746 seconds. This configuration's balance of accuracy and speed underscores its suitability for applications requiring high precision without incurring significant computational costs. MinMax normalization effectively scales the features within a bounded range, while KNN Imputation retains critical data relationships, contributing to exceptional performance.

Another standout combination featured KNN Imputation, Label Encoder, Z-Score normalization, and Decision Tree, achieving a close testing accuracy of 95.96% with a faster testing time of 0.0633 seconds. The slightly lower accuracy compared to MinMax normalization reflects the impact of data scaling choices. Z-Score normalization standardizes features relative to their variance, which may not always align optimally with the underlying data distribution.

Interestingly, iterative imputation, combined with FastText, Robust Scaling, and decision trees, maintained the strong performance of earlier experiments. This configuration achieved a testing accuracy of 95.74%, with a longer testing time of 2.2114 seconds due to the computational demands of FastText. While not the most efficient, this combination demonstrates the robustness of Iterative Imputation and the semantic richness provided by FastText, which improves generalization at the cost of computational efficiency.

The top 10 configurations consistently highlighted the Decision Tree as the best-performing machine learning model across various preprocessing combinations. Its interpretability, robustness, and ability to generalize effectively make it the preferred choice for the TSE pipeline. Among the data cleaning methods, KNN Imputation dominated the top configurations, underscoring its ability to preserve critical relationships in the data. Label Encoder appeared most frequently for data transformation, reflecting its computational efficiency and compatibility with the dataset. Finally, Robust Scaling and MinMax normalization emerged as the most effective normalization techniques, excelling in different scenarios.

This experiment demonstrates the importance of selecting complementary preprocessing methods to optimize the performance of the TSE framework. While KNN Imputation, Label Encoder, MinMax normalization, and Decision Tree proved the most effective overall combination, other configurations, such as those utilizing Iterative Imputation and FastText, offer viable alternatives for specific applications. These findings provide a robust foundation for implementing the TSE framework in satellite anomaly classification and highlight the critical role of method selection in achieving superior accuracy and efficiency.

G. MODEL INTERPRETABILITY IN THE TSE FRAMEWORK USING DECISION TREE

Model interpretability is vital to the Trade-Space Exploration (TSE) framework, particularly when leveraging the Decision Tree as the optimal machine-learning model. The Decision Tree delivers exceptional accuracy (96.07%) and provides a transparent and interpretable structure, aligning with the TSE framework's objective of generating actionable insights for satellite anomaly classification. By visualizing decision rules and feature thresholds, the model facilitates understanding

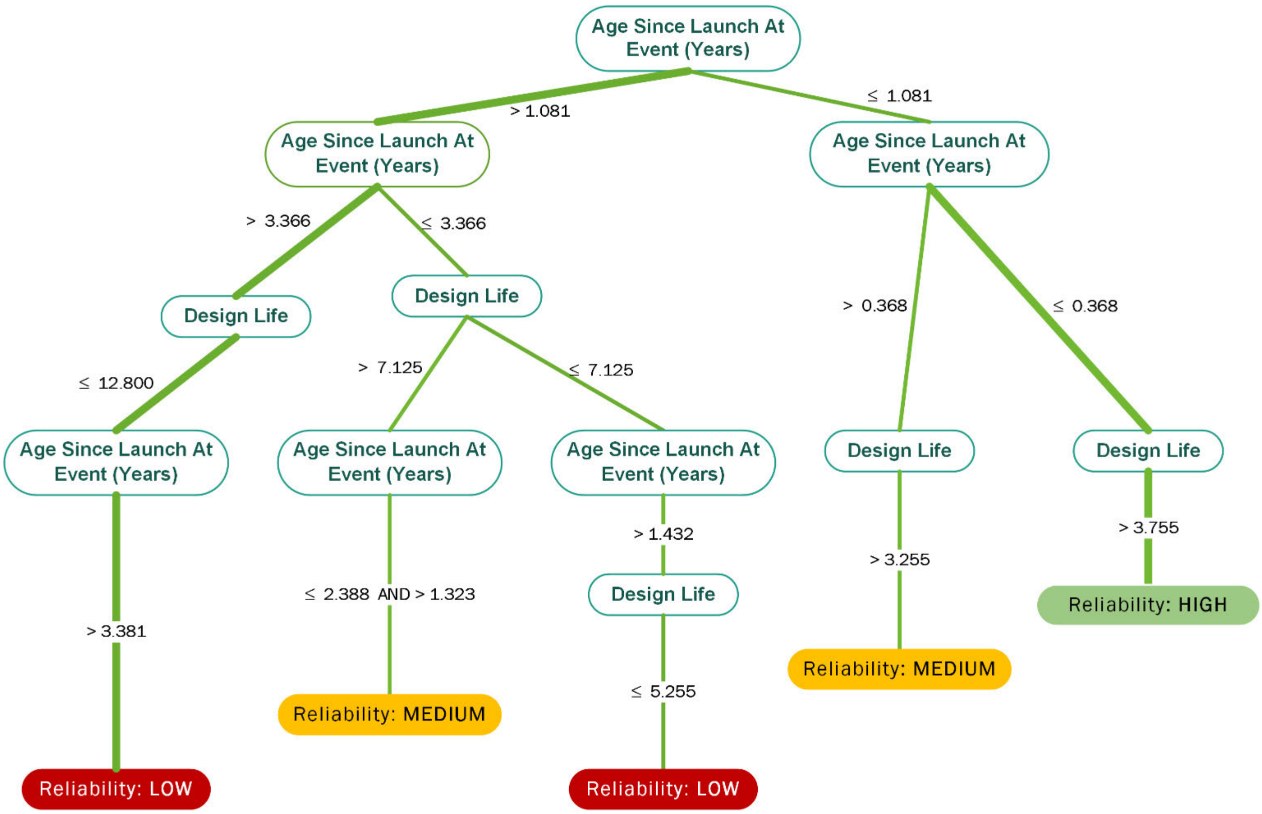

FIGURE 11. Decision tree model for rules with ratio total > 7.97%.

how key attributes—such as Age Since Launch, Design Life, and Orbit Category—influence classification outcomes.

As shown in Figure 11, a simplified Decision Tree serves as an intuitive classifier for satellite anomalies. Age Since Launch at Event is the primary determinant, with satellites older than 1.081 years more likely to have lower reliability. This highlights performance degradation over time, stressing the need for robust design and maintenance. The model further refines classifications using Design Life, predicting that satellites with more than 12.8 years of Design Life are more likely to fall under Medium reliability.

A key pattern observed in the Decision Tree is that satellites older than 3.366 years with a Design Life exceeding 12.8 years are typically categorized under Medium reliability, depending on additional criteria. On the other hand, satellites older than 3.81 years and Design Life less than 12.8 years are more likely to fall under low reliability. Satellites with longer design lives tend to exhibit higher reliability, but the model also highlights critical thresholds where reliability begins to decline. Conversely, satellites classified as High Reliability are typically newer (Age Since Launch ≤ 0.368 years) and have a Design Life greater than 3.755 years. This aligns with advancements in satellite engineering, suggesting that modern designs benefit from improved materials, manufacturing, and operational strategies.

As shown in Fig. 12, the confusion matrix reinforces the model's reliability, with minimal misclassifications between Low, Medium, and High anomaly risk categories. The precision (95.74%) and recall (95.73%) scores indicate a low false positive and false negative rate, ensuring that satellites at risk of anomalies are correctly classified. This is particularly critical for mission planning and risk assessment, as undetected anomalies could lead to mission failures or costly interventions. The high F1 score (95.73%) confirms the model's balanced performance, ensuring that accuracy, precision, and recall are optimally maintained.

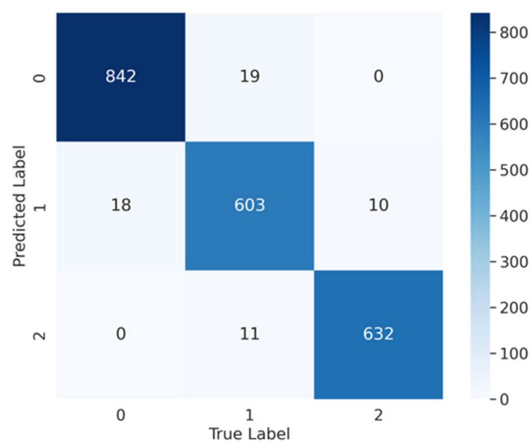

FIGURE 12. Confusion matrix.

From an operational standpoint, the Decision Tree's interpretability allows for rule-based insights that engineers and decision-makers can directly apply. For instance, satellites with an Age Since Launch greater than 1.5 years and a Design Life less than 12 years are consistently classified as Low Reliability, while newer satellites with extended design lifespans are deemed highly reliable. These findings support proactive maintenance strategies and mission design adjustments, ensuring optimized resource allocation for satellite fleets. Additionally, identifying thresholds within Design Life and Event (t) offers guidance on critical failure points, enabling organizations to mitigate operational risks effectively.

Therefore, this Decision Tree analysis strengthens the TSE framework's capability to provide accurate and explainable anomaly classifications. The model ensures informed decision-making in satellite operations by identifying key reliability thresholds, supporting enhanced mission success, and reducing operational failures. Combining high classification performance, low misclassification rates, and strong interpretability makes the Decision Tree an indispensable tool within the TSE-ML pipeline, offering a data-driven foundation for improving satellite reliability in increasingly complex orbital environments.

VI. CONCLUSION

This study introduces the Trade-Space Exploration Machine Learning (TSE-ML) framework, a comprehensive and interpretable pipeline for satellite anomaly classification that addresses the dual challenges of accuracy and computational efficiency. By systematically evaluating 480 configurations across data cleaning, transformation, normalization, and machine learning stages, the framework identified an optimal combination—Iterative Imputation, FastText, Robust Scaling, and Decision Tree—with a high testing accuracy of 95.74%. The Decision Tree model delivered superior classification performance and provided interpretability, revealing critical features such as Age Since Launch, Design Life, and Orbit Category, which drive anomaly classification outcomes. The TSE-ML framework's robust methodology, validated with stratified 5-fold cross-validation, ensures generalizability and practical applicability for satellite operations. It offers actionable insights for improving satellite design, operational planning, and anomaly mitigation, enabling more reliable and efficient systems in increasingly congested orbital environments. By advancing the integration of machine learning into satellite reliability assessment, this work establishes a solid foundation for scalable, transparent, and effective anomaly detection strategies. Future works will focus on extending the TSE-ML framework to real-time telemetry data, integrating additional anomaly types, and exploring its application to broader aerospace systems to further enhance satellite reliability and operational resilience.

REFERENCES

[1] T. Tetik, G. S. Daş, and B. Birgoren, "A multi-objective perspective to satellite design and reliability optimization," *Expert Syst. Appl.*, vol. 246, Jul. 2024, Art. no. 123178, doi: 10.1016/j.eswa.2024.123178.

[2] K. Pollpeter. *China's Role in Making Outer Space More Congested, Contested, and Competitive*. Accessed: Oct. 2021. [Online]. Available: https://www.linkedin.com/company/11049011

[3] R. Holli. (2025). *Catalog of Earth Satellite Orbits*. Accessed: Jan. 14, 2025. [Online]. Available: https://earthobservatory.nasa.gov/features/OrbitsCatalog

[4] R. Mital, K. Cates, J. Coughlin, and G. Ganji, "A machine learning approach to modeling satellite behavior," in *Proc. IEEE Int. Conf. Space Mission Challenges Inf. Technol. (SMC-IT)*, Jul. 2019, pp. 62–69, doi: 10.1109/SMC-IT.2019.00013.

[5] J. C. Green, R. Quinn, T. P. O'Brien, J. Likar, Y. Shprits, A. J. Boyd, S. G. Claudepierre, S. Huston, P. Whelan, and N. Reker, "Developing a comprehensive application for satellite anomaly analysis and attribution," in *Proc. Adv. Maui Org. Space Surveill. Technol. Conf. (AMOS)*, 2020, pp. 1–7.

[6] N. B. A. Rahim and T. L. J. Ferris, "A method to establish a trade-space of system requirements and life cycle cost," *IEEE Syst. J.*, vol. 14, no. 1, pp. 1257–1264, Mar. 2020.

[7] J. Daniels, C. J. Turner, J. Wagner, N. Masoudi, M. Agyemang, G. Hartman, D. Rizzo, D. Gorsich, A. Skowronska, and R. Agusti, "Designing the design space: Evaluating best practices in tradespace exploration, analysis and decision-making," *SAE Int. J. Adv. Current Practices Mobility*, vol. 4, no. 4, pp. 1344–1359, Mar. 2022.

[8] P. A. Oche, G. A. Ewa, and N. Ibekwe, "Applications and challenges of artificial intelligence in space missions," *IEEE Access*, vol. 12, pp. 44481–44509, 2024, doi: 10.1109/ACCESS.2021.3132500.

[9] T. M. Grile and R. A. Bettinger, "Statistical reliability estimation for satellites operating from 1991-2020 with payload reliability focus," in *Proc. 6th Int. Conf. Syst. Rel. Saf. (ICSRS)*, Nov. 2022, pp. 378–386.

[10] N. Ahmad and S. Widyanto, "Detection of satellite anomalies using the dst index," *AIP Conf. Proc.*, vol. 2981, Jan. 2023, Art. no. 040001.

[11] N. A. Rahim, N. S. A. Rahman, S. Abdoli, S. J. Rao, and M. I. M. Mokhtar, "Analysis of failure frequency and failure rate of RF/Antenna subsystems for an Earth station system," *Neuro Quantol.*, vol. 20, no. 10, pp. 8654–8665, 2022.

[12] L. Zhang, J. Yu, D. Tang, D. Han, L. Tian, and J. Dai, "Anomaly detection for spacecraft using hierarchical agglomerative clustering based on maximal information coefficient," in *Proc. 15th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Nov. 2020, pp. 1848–1853.

[13] N. Shazana, A. Rahman, and N. A. Rahim, "Analysis of reliability prediction and maintainability activities of an Earth station system using parallel configurations," in *Proc. Int. Conf. Ind. Eng. Appl.*, vol. 9, 2022, pp. 1647–1655.

[14] H. Elizabeth and M. Tariq. (2025). *Russia Says Its Luna-25 Lunar Lander Has Crashed Into the Moon*. Accessed: Jan. 14, 2025. [Online]. Available: https://www.space.com/russia-says-luna-25-crashes-into-moon

[15] N. T. Rincy and R. Gupta, "A survey on machine learning approaches and its techniques," in *Proc. IEEE Int. Students' Conf. Elect., Electron. Comput. Sci. (SCEECS)*, Feb. 2020, pp. 1–6, doi: 10.1109/SCEECS48394.2020.190.

[16] A. Occhipinti, L. Rogers, and C. Angione, "A pipeline and comparative study of 12 machine learning models for text classification," *Expert Syst. Appl.*, vol. 201, Sep. 2022, Art. no. 117193.

[17] N. A. Rahim and N. A. Nordin, "Reliability model and proposed maintainability activities of earth station system," *Int. J. Electr., Electron. Data Commun.*, vol. 8, no. 5, pp. 20–24, 2020.

[18] A. Juneja and N. N. Das, "Big data quality framework: Pre-processing data in weather monitoring application," in *Proc. Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput. (COMITCon)*, Feb. 2019, pp. 559–563, doi: 10.1109/COMITCon.2019.8862267.

[19] J. Lu, A. Hales, D. Rew, and M. Keech, "Timeline and episode-structured clinical data: Pre-processing for data mining and analytics," in *Proc. IEEE 32nd Int. Conf. Data Eng. Workshops (ICDEW)*, May 2016, pp. 64–67, doi: 10.1109/ICDEW.2016.7495618.

[20] S. J. Hadeed, M. K. O'Rourke, J. L. Burgess, R. B. Harris, and R. A. Canales, "Imputation methods for addressing missing data in short-term monitoring of air pollutants," *Sci. Total Environ.*, vol. 730, Aug. 2020, Art. no. 139140, doi: 10.1016/j.scitotenv.2020.139140.

[21] W. Hameed and N. Ali, "Missing value imputation techniques: A survey," *UHD J. Sci. Technol.*, vol. 7, pp. 72–81, Jul. 2023.

[22] M. Kuhn and K. Johnson, "Data pre-processing," in *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013, pp. 27–59.

[23] M. Abdelaal, C. Hammacher, and H. Schoening, "REIN: A comprehensive benchmark framework for data cleaning methods in ML pipelines," 2023, *arXiv:2302.04702*.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. J. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, Jan. 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[25] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big Data*, vol. 8, no. 1, p. 140, Oct. 2021, doi: 10.1186/s40537-021-00516-9.

[26] C. Guo, W. Yang, C. Liu, and Z. Li, "Iterative missing value imputation based on feature importance," *Knowl. Inf. Syst.*, vol. 66, no. 10, pp. 6387–6414, Oct. 2024, doi: 10.1007/s10115-024-02159-7.

[27] E. Poslavskaya and A. Korolev, "Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding?" 2023, *arXiv:2312.16930*.

[28] X. Rong, "Word2vec parameter learning explained," 2014, *arXiv:1411.2738*.

[29] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Workshop ICLR*, 2013, pp. 1–12.

[30] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "FastText.Zip: Compressing text classification models," 2016, *arXiv:1612.03651*.

[31] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016, *arXiv:1607.01759*.

[32] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," 2017, *arXiv:1712.09405*.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter*, Minneapolis, Minnesota, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[34] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Appl. Soft Comput.*, vol. 133, Jan. 2023, Art. no. 109924, doi: 10.1016/j.asoc.2022.109924.

[35] S. G. K. Patro and K. K. Sahu, "Normalization: A preprocessing stage," *Int. Adv. Res. J. Sci., Eng. Technol.*, vol. 2, no. 3, pp. 20–22, Mar. 2015, doi: 10.17148/iarjset.2015.2305.

[36] Z. Yang, "A modified family of power transformations," *Econ. Lett.*, vol. 92, no. 1, pp. 14–19, Jul. 2006.

[37] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000. [Online]. Available: http://www.jstor.org/stable/2673623

[38] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992, doi: 10.18653/v1/d19-1410.

[39] A. Mutholib, N. A. Rahim, T. S. Gunawan, and A. A. Ahmarofi, "Performance comparison of data preprocessing methods for trade-space exploration with AI model: Case study of satellite anomalies detection," in *Proc. IEEE 10th Int. Conf. Smart Instrum., Meas. Appl. (ICSIMA)*, Jul. 2024, pp. 271–275, doi: 10.1109/ICSIMA62563.2024.10675571.

ABDUL MUTHOLIB (Student Member, IEEE) received the bachelor's degree in informatics (S.Kom) from UIN Syarif Hidayatullah Jakarta, in 2007, and the master's degree in computer and information engineering from International Islamic University Malaysia (IIUM), in 2014, where he is currently pursuing the Ph.D. degree. He is a Lecturer with the Faculty of Science and Technology, UIN Syarif Hidayatullah Jakarta. His research interests include software engineering, image and video processing, and machine learning, with a focus on developing advanced computational techniques to enhance data-driven insights and applications. He is an Active Member of IEEE.

NADIRAH ABDUL RAHIM (Member, IEEE) received the B.Hons. and master's degrees by research in communications engineering from International Islamic University Malaysia (IIUM), Malaysia, in 2007 and 2013, respectively, and the Ph.D. degree in systems engineering from the School of Engineering, University of South Australia, Mawson Lakes, Adelaide, Australia, in June 2018. Previously, she was a Network Integration Engineer with Malaysian Telecommunication Company, Sapura Secured Technologies. She is currently an Assistant Professor with International Islamic University Malaysia, under the Department of Electrical and Computer Engineering (ECE), teaching C++ and Python at the bachelor's level. She is an Engineering Industrial Coordinator for the ECE Students, helping them find internship placements. Apart from that, she is also a student advisor for the ECE Department. Her research interests include satellite communication systems, programming and simulation, artificial intelligence, and engineering management. She is a member of BEM and INCOSE.

TEDDY SURYA GUNAWAN (Senior Member, IEEE) received the B.Eng. degree (cum laude) in electrical engineering from Institut Teknologi Bandung (ITB), Indonesia, in 1998, the M.Eng. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2001, and the Ph.D. degree from the School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia, in 2007. He has been a Professor, since 2019. He was a Visiting Research Fellow with UNSW, from 2010 to 2021, and an Adjunct Professor with Telkom University, from 2022 to 2023. He was the Head of Department with the Department of Electrical and Computer Engineering, from 2015 to 2016, and the Head of Program Accreditation and Quality Assurance for Faculty of Engineering, from 2017 to 2018, International Islamic University Malaysia. He has been a Chartered Engineer (IET), U.K., since 2016, and an Insinyur Profesional Utama (PII), Indonesia, since 2021, a registered ASEAN Engineer, since 2018, an ASEAN Chartered Professional Engineer, since 2020, and an APEC Engineer, since 2023. His research interests include speech and audio processing, biomedical signal processing and instrumentation, image and video processing, and parallel computing. He was awarded the Best Researcher Award from IIUM, in 2018. He was the Chairperson of the IEEE Instrumentation and Measurement Society–Malaysia Section, in 2013, 2014, 2020, and 2021.

MIRA KARTIWI (Member, IEEE) is currently a Professor with the Department of Information Systems, Kulliyyah of Information and Communication Technology, and the Director of the Professional Development, International Islamic University Malaysia (IIUM). She is an experienced Consultant specializing in the health, financial, and manufacturing sectors. Her research interests include health informatics, e-commerce, data mining, information systems strategy, business process improvement, product development, marketing, delivery strategy, workshop facilitation, training, and communications. She was one of the recipients of the Australia Postgraduate Award (APA), in 2004. She was awarded the Higher Degree Research Award for Excellence, in 2007, for her research achievement. She has also been appointed as an editorial board member in local and international journals to acknowledge her expertise.