Draft version June 30, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Identifying Anomalous Asteroids via Predictive Modeling of Physical and Spin Properties based on Orbit and Age

Denario[1]

[1] *Anthropic, Gemini & OpenAI servers. Planet Earth.*

ABSTRACT

Understanding the diverse evolutionary paths of asteroids and identifying objects that deviate from typical trends is crucial for planetary science. Physical and spin properties, such as diameter, spin period, and obliquity, are shaped by complex processes including collisions, thermal radiation forces like the YORP effect, and internal structure, which are not fully determined by current orbital elements and age alone. This study presents an anomaly detection framework to identify asteroids whose observed properties deviate significantly from expected values predicted by their orbit and age. We utilized a large dataset of asteroid properties, including orbital elements (semimajor axis, eccentricity, inclination), estimated age, diameter, spin period, and obliquity. After extensive data preprocessing to handle sparsity, apply logarithmic transformations, and scale features, we trained both Gaussian Process Regression and Neural Network models to predict diameter, spin period, and obliquity from the orbital elements and age. Anomalies were identified by calculating standardized residuals from the GPR models and z-scores of residuals from the NN models, flagging objects whose absolute scores exceeded a predefined threshold. Applying this method identified over 1,100 unique anomalous asteroids. Characterization of this population revealed that these outliers are predominantly larger bodies located on remarkably stable, low-inclination, low-eccentricity orbits within the main belt, and frequently exhibit extreme spin periods that defy typical predictions. These findings suggest that the identified anomalous asteroids likely constitute a physically distinct population, potentially representing primordial planetesimals or objects whose evolution has been governed by unusual events or internal structures, providing valuable targets for further investigation into Solar System formation and evolution.

Keywords: Neural networks, Astrostatistics distributions, Orbital elements, Regression, Outlier detection

1. INTRODUCTION

Asteroids, the rocky and icy remnants left over from the epoch of planet formation, serve as critical witnesses to the physical and chemical conditions of the early Solar System and the subsequent evolutionary processes that have shaped it. Their observable physical and spin properties, including size (diameter), rotation rate (spin period), and the orientation of their spin axis (obliquity), are not static but evolve over billions of years. This evolution is driven by a complex interplay of factors: the initial conditions during accretion, gravitational interactions that influence their orbits, collisions (both catastrophic and non-disruptive) that can alter size and spin state, and non-gravitational forces, such as the Yarkovsky and YORP effects, which subtly change orbits and spins due to thermal radiation. Understanding the current distribution of these properties across the asteroid population provides fundamental insights into the history and ongoing dynamics of the Solar System.

While an asteroid's current orbital elements (semimajor axis, eccentricity, inclination) and its estimated age provide significant constraints on its evolutionary trajectory, they do not fully determine its physical and spin state. This is because several key evolutionary processes are either stochastic (like collisions) or depend on properties not fully captured by orbit and age alone (like the YORP effect, which is highly sensitive to an asteroid's specific shape, surface thermal properties, and internal structure). These complexities introduce substantial variability in physical and spin properties, making it challenging to predict them precisely based solely on orbital parameters and age. Consequently, identifying asteroids whose observed properties deviate significantly

from these expected ranges is difficult but crucial. Such "anomalous" objects may represent populations with unusual formation histories, unique internal compositions or structures, or those that have experienced exceptional evolutionary events, offering valuable clues that lie outside typical evolutionary narratives.

This study addresses this challenge by developing a data-driven anomaly detection framework designed to identify asteroids whose observed physical and spin properties are outliers relative to predictions based on their orbital elements and age. We leverage a comprehensive dataset encompassing asteroid orbital parameters, estimated age, and measured physical and spin properties (diameter, spin period, obliquity). Recognizing the complex, non-linear relationships between these variables and the inherent noise and missingness in observational data, we employ advanced regression techniques. Specifically, we train Gaussian Process Regression (GPR) and Neural Network (NN) models to predict the expected diameter, spin period, and obliquity of an asteroid given its semimajor axis, eccentricity, inclination, and age. GPR models provide probabilistic predictions with associated uncertainties, allowing us to quantify the significance of deviations, while NN models offer robust non-linear function approximation capabilities.

Our method for identifying anomalous asteroids involves quantifying the discrepancy between the observed property value and the value predicted by the models. For GPR models, we calculate standardized residuals, taking into account the model's predictive standard deviation to assess how many standard deviations the observation is from the prediction. For NN models, we calculate the z-score of the residuals to standardize the magnitude of deviations. Asteroids with absolute scores exceeding a predefined threshold are flagged as potential anomalies for the respective property. The final step involves a detailed characterization of the population of identified anomalous asteroids. By analyzing their collective properties, such as their orbital distributions, spectral types, and family memberships, we aim to uncover common traits among these outliers and infer potential physical mechanisms or historical events that could explain their anomalous state. This characterization provides critical context for the identified anomalies, highlighting objects that warrant further targeted observational and theoretical investigation, thereby contributing to a more nuanced understanding of asteroid evolution and the history of the Solar System.

2. METHODS

This study employed a data-driven approach to identify asteroids with physical and spin properties anomalous to their orbital characteristics and estimated age. The methodology involved several distinct stages: data aggregation and preprocessing, exploratory data analysis and feature transformation, predictive modeling using Gaussian Process Regression and Neural Networks, and finally, anomaly identification and characterization of the resulting population.

2.1. *Data Collection and Preprocessing*

The foundation of this study is a comprehensive dataset compiled from multiple sources, providing various observed and derived properties for a large population of asteroids. The initial data were provided as 12 separate CSV files, each containing asteroid identifiers in the first column and a specific property (such as diameter, semimajor_axis, spin_period, age, etc.) in the second.

The first step involved aggregating these disparate files into a single, unified dataset. Each CSV file was loaded into a pandas DataFrame, and columns were immediately renamed to `asteroid_id` and a descriptive property name (e.g., `diameter`, `semimajor_axis`, `age`) for clarity and ease of merging. A master DataFrame was constructed by starting with one property DataFrame and sequentially merging the others using the `asteroid_id` column as the key. An `outer` join strategy was employed for all merges to ensure that every asteroid present in at least one original file was included in the master dataset, preserving information even for objects with missing data for specific properties. The resulting `master_asteroid_dataset.csv` served as the central repository for all subsequent analyses.

2.2. *Exploratory Data Analysis and Feature Engineering*

Prior to developing predictive models, an extensive Exploratory Data Analysis (EDA) was conducted on the merged dataset to understand its characteristics, including data completeness, distributions of variables, and potential issues like skewness and high cardinality. Descriptive statistics revealed significant missing data for several key properties, notably `spin_period` and `obliquity`, as summarized in Table 1. Furthermore, numerical features like `diameter`, `semimajor_axis`, and `spin_period` exhibited strong positive skewness, indicating non-Gaussian distributions. Categorical features, such as `type` and `family`, presented varying levels of cardinality, with `family` having over 1100 unique values.

Based on the EDA findings, a targeted data cleaning and transformation process was implemented to prepare the data for regression modeling. Recognizing that

the availability of target variables differed significantly, three separate preprocessed datasets were created, one for each prediction task (`diameter`, `spin_period`, and `obliquity`). For each task-specific dataset, rows containing missing values for the target variable or any of the designated predictor variables (`semimajor_axis`, `eccentricity`, `inclination`, `age`, and `type`) were removed. This ensured that each modeling task operated on a complete set of relevant features and targets.

To address the observed skewness and potentially linearize relationships for the models, natural logarithm transformations ($\log(x)$) were applied to the `diameter`, `spin_period`, and `semimajor_axis` columns. The `obliquity` target variable, which showed a more symmetric distribution, was not transformed.

Categorical features were handled based on their cardinality. The `type` column, with its manageable number of unique categories, was transformed using one-hot encoding, creating binary indicator columns for each asteroid type. The `family` column, due to its extremely high cardinality, was deemed unsuitable for direct use as a predictor in this modeling phase and was excluded from the feature set to avoid dimensionality issues.

Finally, all numerical predictor variables (`semimajor_axis` (log-transformed), `eccentricity`, `inclination`, and `age`) were standardized using `scikit-learn`'s `StandardScaler`. This process scales each feature to have a mean of zero and a standard deviation of one, which is crucial for models sensitive to feature scales, such as Neural Networks and Gaussian Processes. The scaler was fitted exclusively on the training portion of the data for each target variable dataset and then applied to both the training and testing sets to prevent data leakage. The three resulting, fully preprocessed datasets (one for each target property prediction) were saved for use in the modeling stage.

2.3. *Predictive Modeling*

To predict the expected values of diameter, spin period, and obliquity based on orbital elements and age, two distinct regression modeling techniques were employed: Gaussian Process Regression (GPR) and Neural Networks (NN). This dual approach allowed for comparison between a probabilistic model that provides uncertainty estimates (GPR) and a powerful non-linear function approximator (NN). For each of the three target variables, the corresponding preprocessed dataset was split into an 80% training set and a 20% testing set.

2.3.1. *Gaussian Process Regression*

Gaussian Process Regression models were chosen for their ability to provide probabilistic predictions, including both a mean prediction and a predictive variance (standard deviation). This uncertainty estimate is particularly valuable for quantifying the significance of deviations in the anomaly detection phase.

For each target variable (log(diameter), log(spin_period), obliquity), a `GaussianProcessRegressor` model was instantiated. A composite kernel was defined to capture the underlying relationships in the data. A Radial Basis Function (RBF) kernel was used to model the smooth, non-linear relationships between the input features and the target variable. This was combined with a WhiteKernel, which accounts for the noise in the observations. The GPR models were trained on the respective training datasets. During training, the model optimizes the hyperparameters of the kernel (such as the length-scale of the RBF kernel and the noise level of the WhiteKernel) by maximizing the log-marginal-likelihood of the data. After training, predictions were generated for the test set, yielding both the predicted mean (μ) and the predictive standard deviation (σ) for each data point. The three trained GPR models were serialized and saved to disk using `joblib` for later use in anomaly identification.

2.3.2. *Neural Network*

Neural Networks, specifically Multi-Layer Perceptrons (MLPs), were employed as robust non-linear regressors capable of learning complex dependencies between the orbital/age features and the physical/spin properties.

For each target variable, an MLP architecture was defined. The input layer's size matched the number of features in the preprocessed data (scaled numerical features plus one-hot encoded type features). The network consisted of three hidden layers with 128, 64, and 32 neurons, respectively, each using the Rectified Linear Unit (ReLU) activation function. A dropout layer with a rate of 0.2 was included after each hidden layer to regularize the model and mitigate overfitting. The output layer consisted of a single neuron with a linear activation function, suitable for regression tasks.

The models were compiled using the Adam optimizer and the Mean Squared Error (MSE) as the loss function, aiming to minimize the average squared difference between predictions and observed values. Given that optimal NN hyperparameters are often data-dependent, a hyperparameter tuning process was necessary. A search was conducted using frameworks like `scikit-learn`'s `GridSearchCV` or `Hyperopt` to explore different combinations of hyperparameters (e.g., learning rate, network structure variations, dropout rates). This computationally intensive tuning process was parallelized across 128 CPU cores. The best-performing model configuration for each target variable was then trained on its respec-

tive training set. During training, a validation split (e.g., 20% of the training data) was used, and an early stopping callback monitored the validation loss to halt training when performance improvements plateaued, further preventing overfitting and determining the optimal number of training epochs. The architecture and learned weights of the three final, trained NN models were saved.

2.4. *Anomaly Identification and Characterization*

With the trained GPR and NN models for each target property, the final stage involved identifying asteroids whose observed properties deviate significantly from the model predictions and characterizing this anomalous population. Predictions were generated for every asteroid in the complete, preprocessed datasets (combining training and testing partitions) using all six trained models (three GPR and three NN).

Anomaly scores were calculated differently for the two model types to leverage their respective outputs.

- For the GPR models, which provide predictive uncertainty, the anomaly score for asteroid i and a given target variable was defined as the standardized residual: $S_{GPR,i} = (y_i - \mu_i)/\sigma_i$, where y_i is the observed (and potentially log-transformed) value, μ_i is the GPR's predicted mean, and σ_i is the GPR's predictive standard deviation. This score represents how many predictive standard deviations the observed value is away from the predicted mean.
- For the NN models, which provide only point predictions, the raw residual was first calculated: $r_{NN,i} = y_i - \hat{y}_i$, where $\hat{y}_i$ is the NN's prediction. To standardize these residuals and make their magnitude comparable across different properties, the z-score of the residuals was computed: $S_{NN,i} = (r_{NN,i} - \text{mean}(r_{NN}))/\text{std}(r_{NN})$.

An asteroid was flagged as anomalous for a specific property and model if the absolute value of its calculated anomaly score exceeded a predefined threshold. A threshold of $|S_i| > 3$ was used, corresponding to deviations greater than three standard deviations from the predicted value (either predictive standard deviation for GPR or residual standard deviation for NN). This process generated six boolean flags for each asteroid, indicating potential anomaly status for each model-property combination.

These anomaly scores and flags were merged back into the original `master_asteroid_dataset.csv` using the `asteroid_id`, creating a comprehensive results table that included all initial data alongside the calculated anomaly scores and flags.

Finally, the population of identified anomalous asteroids was characterized. Asteroids flagged as anomalous by at least one of the models were isolated. A comparative statistical summary was generated for this anomalous population, contrasting their properties (orbital elements, diameter, spin period, obliquity, age) with those of the non-anomalous majority. The distributions of their orbital elements (`semimajor_axis`, `eccentricity`, `inclination`), spectral types (`type`), and family memberships (`family`) were specifically analyzed to identify common characteristics or biases within the outlier group. This characterization aimed to infer potential physical mechanisms, formation pathways, or evolutionary events that might explain their anomalous state relative to predictions based solely on orbit and age, thereby highlighting promising targets for future investigation. The results of this comparative analysis were saved to a text file.

3. RESULTS

The results of this study detail the process of identifying asteroids with physical and spin properties that are anomalous relative to predictions based on their orbital elements and estimated age. This section presents the outcomes of data preparation, predictive modeling using Gaussian Process Regression (GPR) and Neural Networks (NN), and the subsequent characterization of the identified anomalous population.

3.1. *Data aggregation and preprocessing*

The initial step involved the aggregation of data from 12 source files into a single master dataset comprising over 1.7 million asteroid entries. This aggregation process, as described in the Methods, utilized an outer join to preserve data for all asteroids present in any source file. Initial exploratory data analysis revealed significant data sparsity, particularly for physical and spin properties. While orbital elements such as semimajor axis, eccentricity, and inclination were available for a large fraction of the catalog, spin_period data were present for only approximately 3.2% of entries, and obliquity data for a mere 0.2%. The extent of missing data across various features is visualized in the heatmap shown in Figure 1. The significant sparsity for target variables necessitated the creation of three distinct, cleaned datasets, one for each target variable (diameter, spin_period, and obliquity), by removing rows with missing values in either the target or any of the predictor variables (semimajor axis, eccentricity, inclination, age, type).

The resulting cleaned datasets used for modeling had varying sizes: 10,340 asteroids for the diameter prediction task, 6,396 for the spin_period task, and 1,626

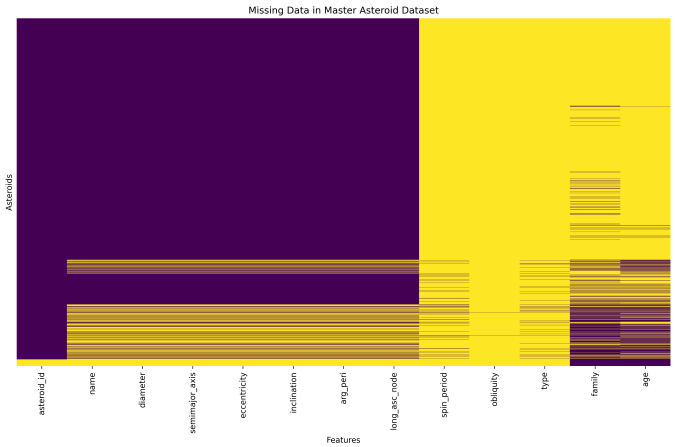

Figure 1. Missing data heatmap for the master asteroid dataset. Yellow indicates data presence, purple absence. Orbital parameters are largely complete, whereas physical properties like spin period and obliquity are highly sparse, necessitating the creation of feature-specific datasets for modeling.

for the obliquity task. Prior to modeling, distributions of numerical features were examined. Strong positive skewness was observed in diameter, spin_period, and semimajor axis. To mitigate the impact of this skewness and potentially linearize relationships, natural logarithm transformations were applied to these variables. The effectiveness of this transformation was confirmed by examining feature histograms for the datasets, which showed distributions closer to symmetric or Gaussian shapes for the log-transformed variables. Figures 2, 3, and 4 display the distributions of features used for the diameter, spin period, and obliquity models, respectively, highlighting the effect of the log transformation on semimajor axis, diameter, and spin period. The categorical asteroid type feature, with 20-25 unique classes across the cleaned datasets, was one-hot encoded. The high-cardinality family feature was excluded from the modeling process to avoid excessive dimensionality. Finally, all numerical predictor variables were standardized using z-scoring based on the training data mean and standard deviation, a crucial step for models sensitive to feature scaling.

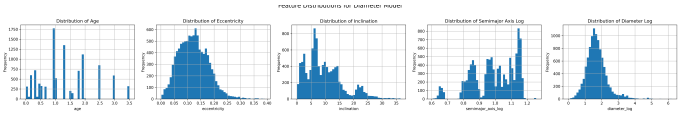

Figure 2. Histograms showing the frequency distribution of features used for the diameter prediction model: age, eccentricity, inclination, log-transformed semimajor axis, and log-transformed diameter. The log transformation improves the symmetry of the distributions for semimajor axis and diameter.

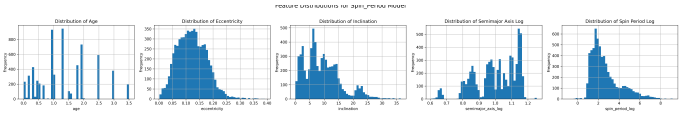

Figure 3. Histograms showing the distributions of features used in the Spin_Period model. The distributions of semimajor axis and spin_period are shown after a natural logarithm transformation to mitigate skewness, while age, eccentricity, and inclination distributions are also displayed. These plots illustrate the characteristics of the data input to the predictive models.

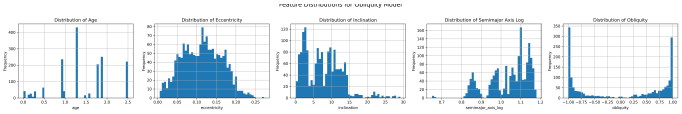

Figure 4. Histograms showing the distribution of features used for training the obliquity prediction model. The features include asteroid age, eccentricity, inclination, log-transformed semimajor axis, and the target variable, obliquity. These plots visualize the characteristics and value ranges of the data subset available for obliquity modeling.

3.2. *Predictive modeling performance*

Predictive models were trained using both Gaussian Process Regression (GPR) and Neural Networks (NN) to estimate the expected values of diameter, spin_period, and obliquity based on the preprocessed orbital elements and age.

3.2.1. *Gaussian process regression performance*

GPR models were trained for each target variable using a composite kernel consisting of a Radial Basis Function (RBF) and a WhiteKernel. The RBF component models the underlying smooth function relating inputs to outputs, while the WhiteKernel accounts for irreducible noise or variance. The kernel hyperparameters, optimized during training by maximizing the log-marginal-likelihood, provide insights into the structure of the data and the model's fit.

For the log(diameter) model, the optimized kernel was $1.41^2 \times \text{RBF}(\text{length\_scale} = 3.35) + \text{WhiteKernel}(\text{noise\_level} = 0.235)$. The relatively small length scale (3.35 in standardized units) suggests the model learned moderately complex, non-linear relationships between the predictors and log(diameter). The noise level of 0.235 indicates a measurable amount of variance in log(diameter) that is not explained by the input features. The predictive performance is illustrated in Figure 5, showing predicted versus true values, and Figure 6, showing residuals versus predicted values.

The log(spin_period) model resulted in a kernel of $1.79^2 \times \text{RBF}(\text{length\_scale} = 100) +$

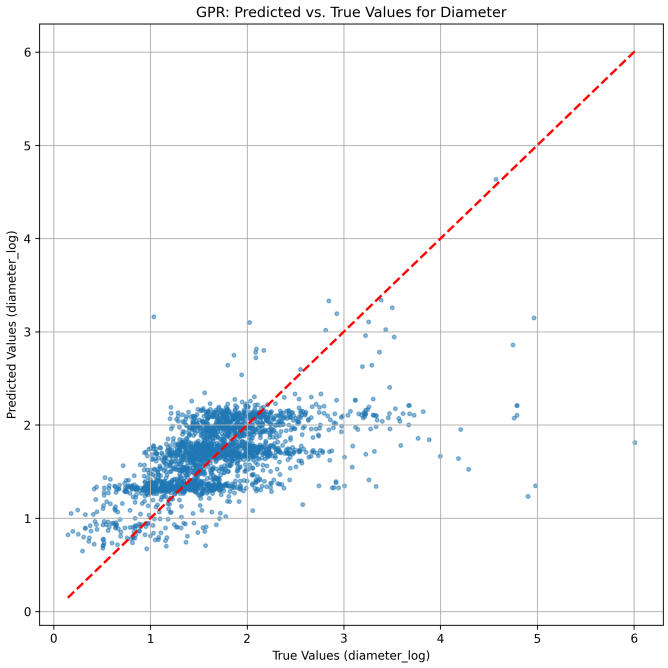

Figure 5. Scatter plot of predicted versus true values for log-transformed asteroid diameter from the Gaussian Process Regression model. The dashed red line represents perfect prediction. The plot shows a positive correlation, indicating the model captures the general relationship, but significant scatter is present, particularly for larger diameters, revealing the model's limitations in predicting precise values for these objects.

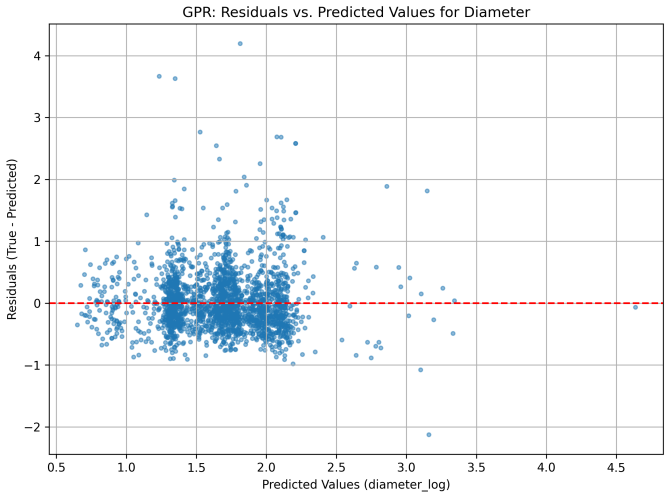

Figure 6. Residuals versus predicted values for the Gaussian Process Regression model predicting log(diameter). The plot shows the distribution of prediction errors, indicating increased scatter for larger predicted diameters.

WhiteKernel(noise_level = 2.07). The extremely large length scale (100) suggests the model found the relationship between predictors and log(spin_period) to be very smooth, almost linear over the range of the data. The significantly higher noise level (2.07) compared to the diameter model indicates that a large proportion of the variance in log(spin_period) is treated as noise by the model, suggesting the input features have limited power in predicting spin_period compared to diameter. Figure 7 shows the predicted versus true values for log(spin_period), where predictions cluster near the mean, reflecting this limited predictive power. Figure 8 shows the distribution of predictive uncertainty for this model.

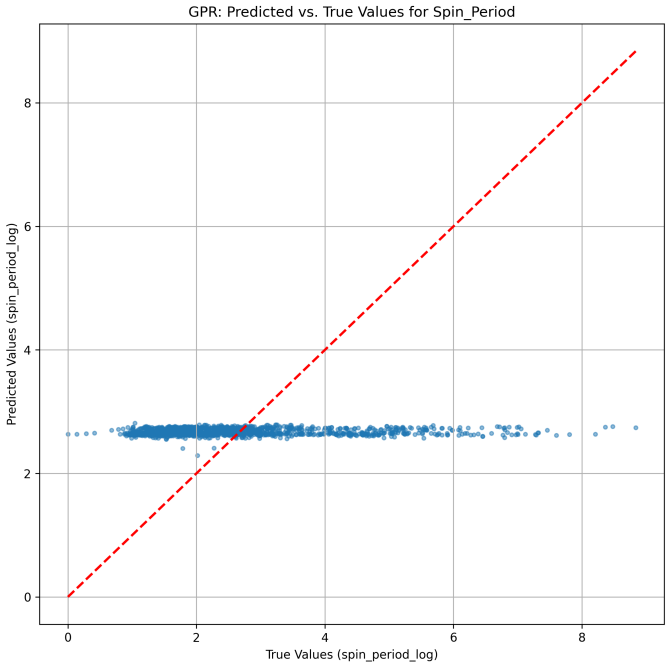

Figure 7. Gaussian Process Regression predicted versus true values for the natural logarithm of asteroid spin period. The clustering of points around a horizontal line indicates that the model's predictions are close to the mean of the true values, demonstrating limited predictive power from the input features for this property.

For the obliquity model, the kernel was $0.851^2 \times \text{RBF}(\text{length\_scale} = 0.042) + \text{WhiteKernel}(\text{noise\_level} = 1e-10)$. The extremely small length scale (0.042) implies the model attempted to fit a highly complex, rapidly varying function. The near-zero noise level is indicative of potential overfitting, where the model attributes almost all observed variance to the complex function learned by the RBF kernel, leaving very little residual variance to be modeled as noise. Figure 9 shows the predicted versus true values, Figure 10 shows the predictive uncertainty distribution, and Figure 11 shows the residuals for the obliquity

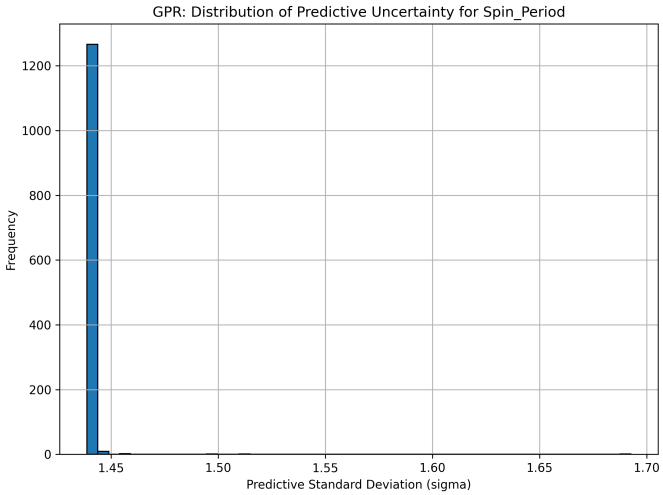

Figure 8. Histogram showing the distribution of predictive uncertainty (standard deviation) from the Gaussian Process Regression model for log-transformed spin_period. The distribution is sharply peaked around a standard deviation of approximately 1.44 (corresponding to the model's noise level), indicating the model makes predictions with consistently low uncertainty despite having limited predictive power for this variable.

GPR model, all reflecting the challenges in predicting this property with the available data.

Visual inspection of predicted-vs-true plots for GPR models showed a positive correlation for diameter (Figure 5), albeit with significant scatter, particularly for larger objects. For spin_period, predictions were largely clustered around the mean (Figure 7), reflecting the model's difficulty in capturing the wide range of observed values, consistent with the high noise level inferred from the kernel and the predictive uncertainty shown in Figure 8. The obliquity model showed a relationship (Figure 9), but predictions were dispersed and residuals structured (Figure 11), indicating limited predictive accuracy. A key advantage of GPR is its ability to provide predictive uncertainty (σ_i), which was utilized in the anomaly scoring.

3.2.2. *Neural network performance*

Multi-Layer Perceptron (MLP) models were trained with three hidden layers and dropout for each target variable. Training progress was monitored using loss curves, showing a decrease in both training and validation loss. Early stopping was employed to prevent overfitting, restoring model weights from the epoch with the best validation loss. The training and validation loss curves for the diameter, spin period, and obliquity NN models are shown in Figures 12, 13, and 14, respectively.

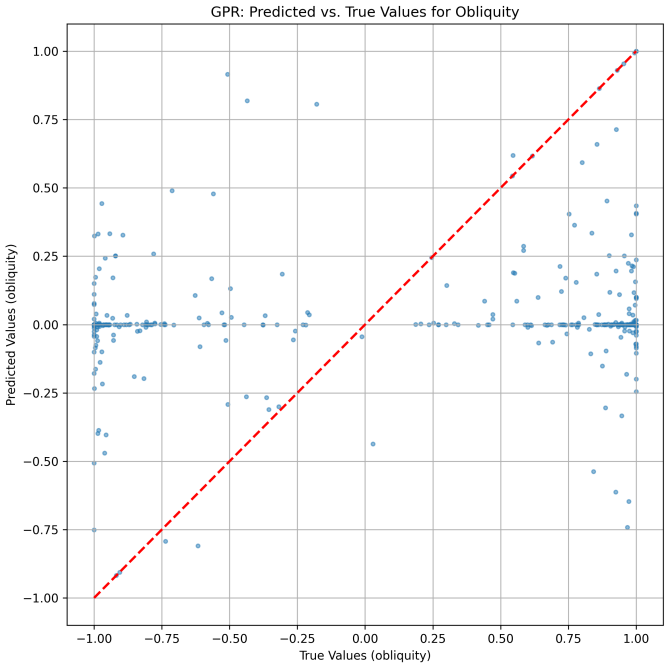

Figure 9. Gaussian Process Regression predicted versus true values for asteroid obliquity. The dashed line indicates perfect prediction. The model shows a dispersed relationship, struggling to predict obliquity across the observed range.

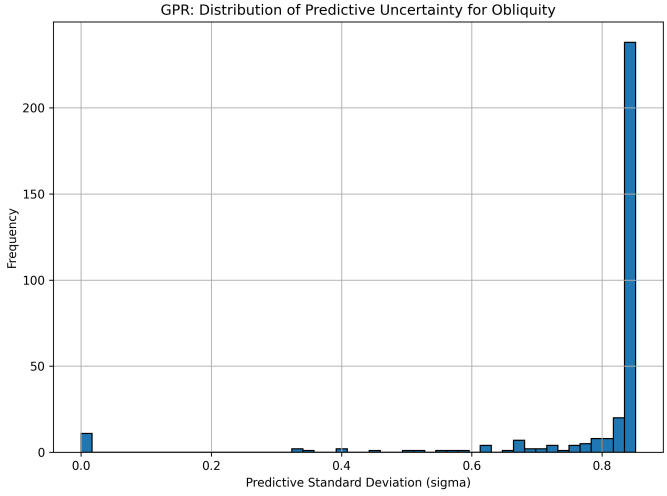

Figure 10. Histogram showing the distribution of predictive standard deviation for the Gaussian Process Regression model predicting asteroid obliquity. The distribution reveals high predictive uncertainty for most asteroids in the dataset.

The predictive performance of the NN models, assessed on the test sets, was broadly comparable to that of the GPR models in terms of mean prediction accuracy. The NN model for diameter showed good correlation between predicted and true values (Figure 15), with

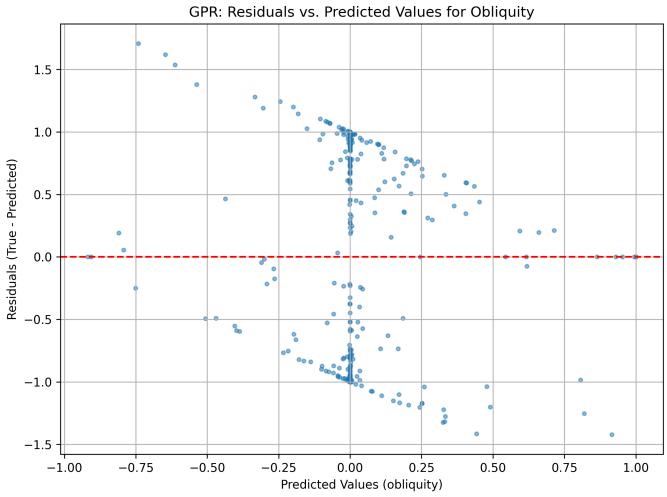

Figure 11. Residuals vs. predicted values for the Gaussian Process Regression model predicting asteroid obliquity. The plot illustrates the model's predictive performance and the distribution of errors, showing structured and dispersed residuals.

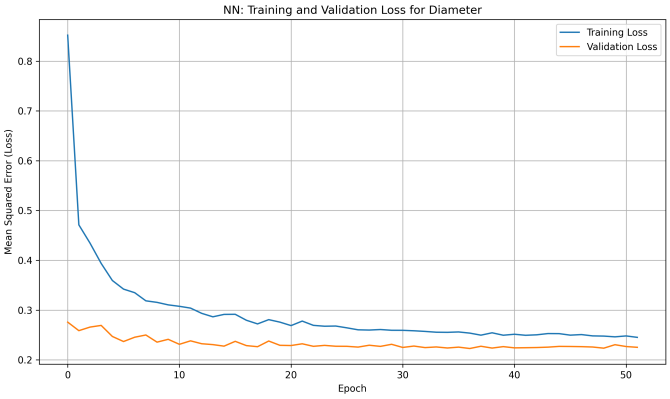

Figure 12. Neural Network training and validation loss curves for the asteroid diameter prediction model. The plot shows the Mean Squared Error decreasing over epochs, with validation loss tracking training loss before plateauing, indicating effective training and prevention of overfitting.

residuals centered around zero but showing increasing scatter for larger predicted values (Figure 16).

Similar to the GPR, the NN model for spin_period also struggled to capture the full range of observed values, producing predictions weakly correlated with the true spin_periods (Figure 17). The large spread of residuals for spin period (Figure 18) further highlights this difficulty.

For obliquity, the NN model also showed limited predictive accuracy, with scattered predictions (Figure 19) and structured residuals (Figure 20).

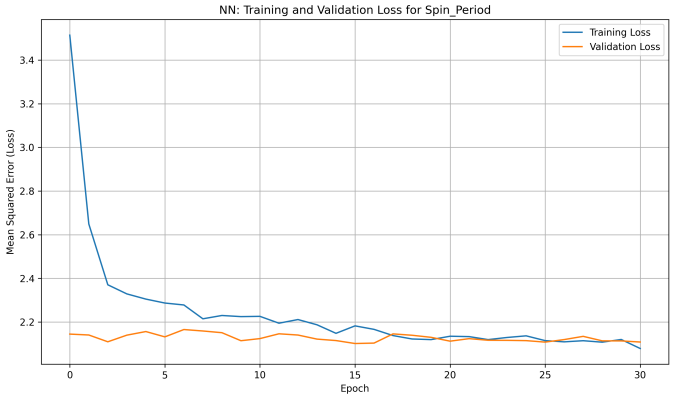

Figure 13. Neural Network training and validation loss for asteroid spin_period prediction. The mean squared error decreases with training epochs, converging to a plateau, demonstrating model learning and generalization.

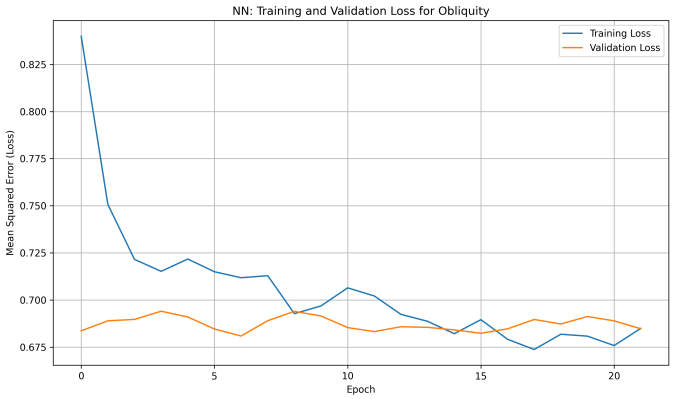

Figure 14. Neural Network training and validation loss curves for the obliquity model, showing Mean Squared Error (Loss) as a function of training epoch. The decrease in training loss and subsequent plateauing of validation loss indicate the model learned from the data, highlighting the challenge in accurately predicting asteroid obliquity.

The final validation losses confirmed these observations: 0.225 for diameter, 2.108 for spin_period, and 0.684 for obliquity, numerically reinforcing that spin_period was the most challenging property to predict given the input features. Both modeling approaches thus independently indicated that orbital elements and age are more predictive of diameter than of spin_period or obliquity.

3.3. *Anomaly identification and characterization*

The core objective was to identify asteroids whose observed properties significantly deviate from the values predicted by the models based on their orbit and age. Anomaly scores were calculated for each asteroid in the respective cleaned datasets using the trained GPR and NN models.

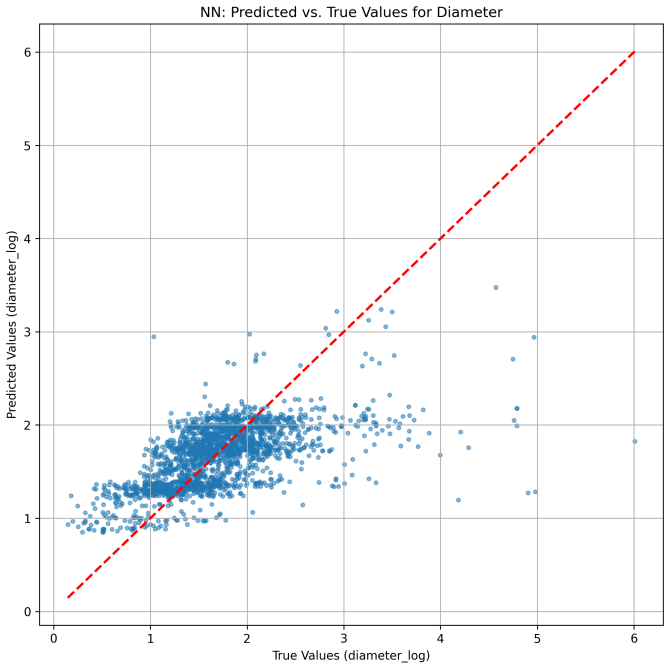

Figure 15. Neural Network predicted versus true values for log-transformed asteroid diameter. The plot shows a positive correlation indicating the model captures the general trend, but significant scatter, particularly at higher values, reveals limitations in predicting larger diameters.

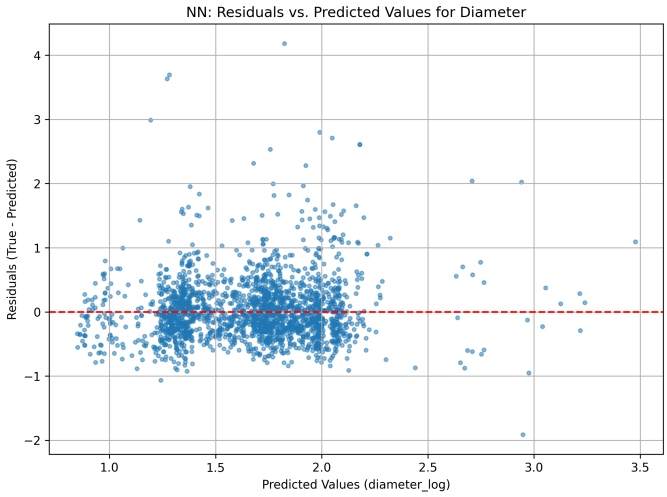

Figure 16. Residuals (True - Predicted) versus predicted values for the log-transformed diameter from the Neural Network model. The residuals are centered around zero, indicating good overall prediction accuracy, but show increasing scatter with higher predicted values, suggesting potential heteroscedasticity. The distribution of residuals informs the model's predictive performance and is used for anomaly detection.

3.3.1. *Anomaly identification*

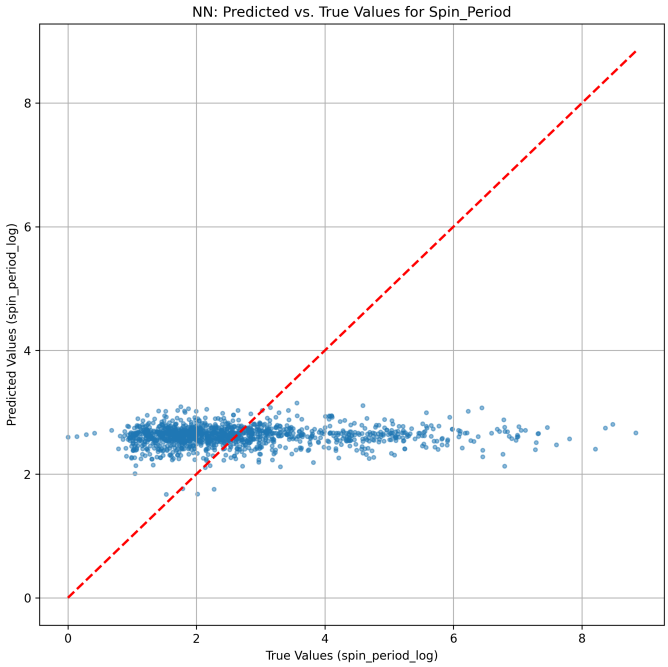

Figure 17. Scatter plot showing Neural Network predicted values versus true values for the natural logarithm of asteroid spin period. The lack of clear correlation and clustering of predictions around the mean indicates that the model struggles to predict spin period from the input features.

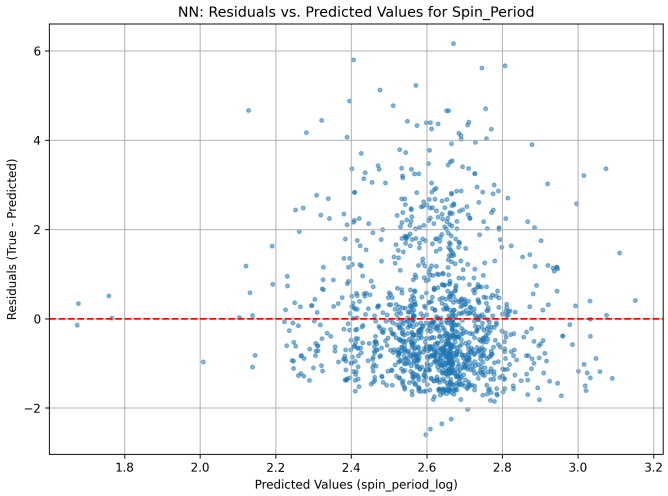

Figure 18. Residuals versus predicted values for the neural network model predicting log(spin_period). The large spread of residuals around zero highlights the model's difficulty in accurately predicting asteroid spin periods from the given features.

For GPR models, the anomaly score for an asteroid i and a given property was defined as the standardized residual, $S_{GPR,i} = (y_i - \mu_i)/\sigma_i$, where y_i is the observed (potentially log-transformed) value, μ_i is the GPR's pre-

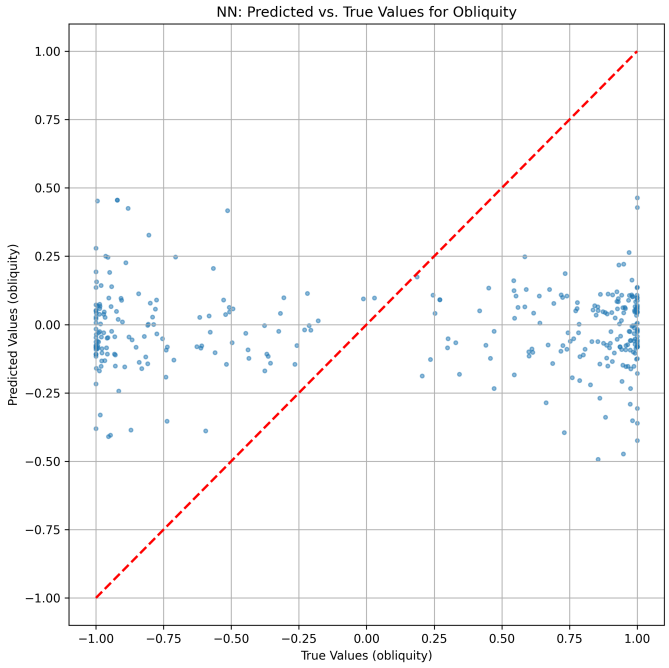

Figure 19. Predicted versus true values for asteroid obliquity using the Neural Network model. The scatter of points around the red dashed line (representing perfect prediction) indicates that the model struggled to accurately predict obliquity, consistent with its low anomaly yield.

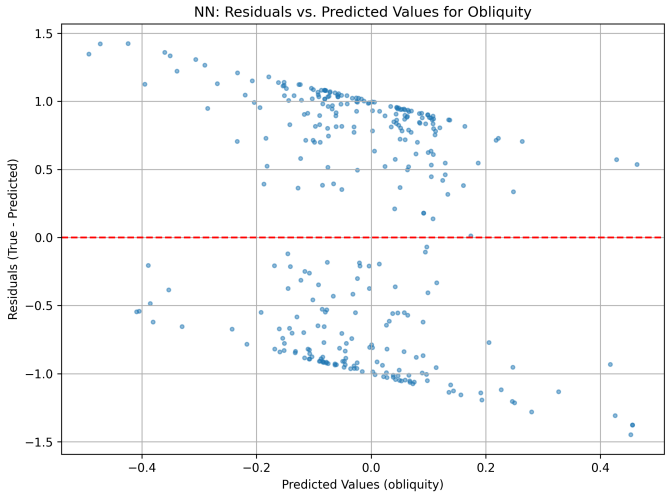

Figure 20. Neural Network model residuals for asteroid obliquity. The plot shows the difference between true and predicted obliquity values against the predicted values. The non-random pattern of residuals indicates the model struggles to accurately predict the full range of obliquity values, consistent with the small dataset size for this property.

dicted mean, and σ_i is the GPR's predictive standard deviation. This score quantifies the deviation in terms of the model's inherent uncertainty. For NN models, the anomaly score was the z-score of the raw residual, $S_{NN,i} = (y_i - \hat{y}_i - \text{mean}(r_{NN}))/\text{std}(r_{NN})$, where $\hat{y}_i$ is the NN prediction and r_{NN} are the residuals across the dataset. This standardizes the residual magnitude.

An asteroid was flagged as anomalous for a specific property if the absolute value of its anomaly score from either the GPR or NN model for that property exceeded a threshold of 3. This threshold corresponds to deviations greater than three standard deviations, indicating a statistically significant outlier. This process generated six potential anomaly flags per asteroid.

Applying this criterion across all models and properties identified a total of **1,138 unique asteroids** flagged as anomalous by at least one model. The distribution of flags across models and properties was as follows:

- GPR Diameter: 973 anomalies
- NN Diameter: 989 anomalies
- GPR Spin_Period: 136 anomalies
- NN Spin_Period: 141 anomalies
- GPR Obliquity: 3 anomalies
- NN Obliquity: 0 anomalies

The majority of flagged anomalies are associated with diameter, reflecting the models' better predictive power for this property, which allows for more confident identification of outliers. The smaller number of spin_period anomalies aligns with the models' difficulty in predicting this property, resulting in larger predictive uncertainties (GPR) or residual variances (NN), which make it harder to exceed the $|S_i| > 3$ threshold. The near-absence of obliquity anomalies is likely due to the very small dataset size available for this property (1,626 asteroids) and the GPR model's potential overfitting, leaving minimal residual variance to identify outliers.

3.3.2. *Profile of the anomalous population*

To understand the nature of these identified anomalies, a detailed comparative statistical analysis was conducted between the population of 1,138 anomalous asteroids and the remaining non-anomalous asteroids in the master dataset. The results of this comparison are summarized in Table 1.

The table reveals a consistent and distinctive profile for the anomalous group. The most striking characteristic is their size. The mean diameter of anomalous asteroids is 63.81 km, which is over 30 times larger than the mean diameter of 1.96 km for the vast majority of non-anomalous asteroids in the dataset. This indicates

Table 1. Comparative Descriptive Statistics of Anomalous vs. Non-Anomalous Asteroids

Feature	Population	Count	Mean	Std Dev	Min	25%	50%	75%	Max
diameter (km)	**Anomalous**	**1,138**	**63.81**	**47.03**	**2.15**	**35.40**	**53.20**	**117.45**	**498.10**
	Non-Anomalous	1,462,706	1.96	11.72	0.00	0.74	1.10	1.64	5909.55
spin_period (hr)	**Anomalous**	**1,136**	**261.73**	**810.62**	**2.53**	**7.03**	**14.21**	**16.51**	**9567.39**
	Non-Anomalous	62,767	73.47	396.93	0.04	5.16	8.96	28.89	10167.60
inclination (deg)	**Anomalous**	**1,138**	**3.13**	**4.61**	**0.24**	**1.01**	**1.51**	**2.21**	**34.92**
	Non-Anomalous	1,462,706	9.29	6.66	0.01	4.33	7.87	12.70	175.98
eccentricity	**Anomalous**	**1,138**	**0.12**	**0.05**	**0.01**	**0.06**	**0.13**	**0.17**	**0.27**
	Non-Anomalous	1,462,706	0.16	0.09	0.00	0.09	0.15	0.20	0.997
age (Gyr)	**Anomalous**	**1,138**	**1.24**	**1.14**	**0.01**	**0.01**	**1.30**	**2.50**	**3.50**
	Non-Anomalous	315,545	1.01	0.87	0.00	0.30	0.93	1.50	3.50

that the models, trained primarily on the distribution of smaller bodies, consistently underpredicted the diameter of these large objects based on their orbital parameters and age.

Regarding spin properties, the anomalous population exhibits extreme spin_periods. While the mean spin_period (261.73 hr) is significantly higher than the non-anomalous mean (73.47 hr), the extremely large standard deviation (810.62 hr) highlights the presence of both exceptionally slow rotators (long periods) and potentially very fast rotators (though the mean is high, the distribution is broad). These extreme spin states are not well-predicted by the models, contributing to their anomalous flags.

The orbital characteristics of the anomalous population are also distinct. Counter-intuitively, these outliers tend to occupy more stable and less excited orbits. Their mean inclination (3.13°) and eccentricity (0.12) are significantly lower than the means for the non-anomalous population (9.29° and 0.16, respectively). This suggests that while their physical and spin properties are unusual according to the models, their orbits are relatively 'typical' or even less dynamically evolved than the average asteroid. They are predominantly located within the main asteroid belt. Figure 21 provides a visual comparison of the distributions for diameter, spin period, inclination, and eccentricity between the anomalous and non-anomalous populations, clearly showing these differences.

The distribution of estimated ages shows that the anomalous population has a slightly higher mean age (1.24 Gyr vs 1.01 Gyr), with a median age similar to the overall population median. However, the age estimates themselves have uncertainties and represent a simplified proxy for a complex evolutionary history.

In summary, the identified anomalous asteroids are primarily characterized by a combination of large size,

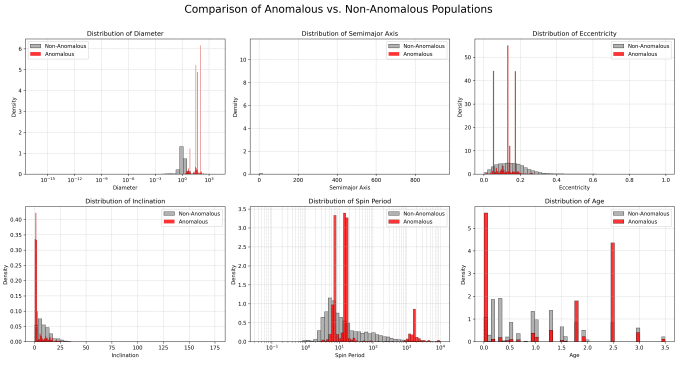

Figure 21. Comparison of property distributions for anomalous (red) and non-anomalous (gray) asteroid populations. Anomalous asteroids are notably larger, occupy more stable orbits (lower inclination and eccentricity), and exhibit different spin_period distributions.

stable orbits, and extreme spin states, as detailed in Table 1 and visualized in Figure 21.

3.4. *Physical interpretation and implications*

The consistent profile of the identified anomalous asteroids suggests they do not represent random noise or errors but rather a physically distinct subpopulation. The combination of large size and stable orbits points towards a group of objects that have experienced a different evolutionary pathway compared to the majority of the observed asteroid population.

The large size of these objects makes them less susceptible to orbital changes driven by non-gravitational forces like the Yarkovsky effect and significantly more resistant to catastrophic disruption from collisions. Their stable, low-inclination, low-eccentricity orbits are characteristic of the primordial main belt, suggesting they may be among the least dynamically evolved objects since the early Solar System. This combination of large size and stable orbits aligns with the hypothesis that these asteroids could represent surviving **primordial**

planetesimals—the building blocks from which the planets formed. Their large mass would have helped them avoid fragmentation and major orbital perturbations over billions of years.

Furthermore, the YORP effect, which significantly alters the spin periods and obliquities of smaller asteroids over time, becomes much less efficient for larger bodies (typically > 40 km). The models used in this study are implicitly trained on the dominant population of smaller asteroids where YORP is active. The large size of the anomalous asteroids means their spin states are less likely to be governed by the same YORP-driven evolution as the smaller bodies. This could explain why their spin periods are often extreme and defy the predictions derived from models that are more applicable to the smaller asteroid population. The observed extreme spin states (both very fast and very slow) could be remnants of their initial formation spin, altered only by rare, large-scale events like significant collisions or internal processes, rather than the gradual YORP spin-up/spin-down experienced by smaller bodies.

Alternative explanations could involve a **distinct collisional history**. While age is included as a predictor, it simplifies the complex impact history. These objects might have experienced fewer catastrophic collisions or a different type of impacts (e.g., grazing impacts or mergers) that altered their size and spin state in ways not typical for the general population, without significantly disrupting their orbits. Differences in **internal structure or composition** could also play a role, affecting both their resilience to collisions and their response to thermal forces, leading to physical properties that deviate from predictions based on external orbital parameters and simplified age.

In summary, the anomaly detection framework successfully identified a population of asteroids characterized by large size, stable orbits, and extreme spin states. These characteristics strongly suggest they represent a physically meaningful group, likely consisting of primordial planetesimals whose evolution has been different from the majority of asteroids due to their size and initial conditions. These objects are high-priority targets for future observational and theoretical studies aimed at understanding the initial conditions and evolutionary pathways of the Solar System.

4. CONCLUSIONS

Asteroids provide a window into the formation and evolution of the Solar System. While their physical and spin properties are shaped by complex processes related to their orbits and age, accurately predicting these properties is challenging due to stochastic events like collisions and non-gravitational forces like the YORP effect. This study addressed the need to identify asteroids whose observed properties deviate significantly from these expected trends, potentially highlighting objects with unusual histories.

We developed a data-driven anomaly detection framework utilizing Gaussian Process Regression (GPR) and Neural Network (NN) models. These models were trained on a comprehensive dataset of asteroid orbital elements (semimajor axis, eccentricity, inclination), estimated age, and observed physical and spin properties (diameter, spin period, obliquity). After extensive data preprocessing, including logarithmic transformations and feature scaling, the models predicted the expected values of diameter, spin period, and obliquity based on the input features. Anomalies were identified by calculating standardized residuals (for GPR) or z-scores of residuals (for NN), flagging asteroids with absolute scores exceeding a threshold of 3.

Applying this framework led to the identification of 1,138 unique asteroids flagged as anomalous by at least one model. The majority of anomalies were identified based on diameter, reflecting the models' relatively better predictive power for this property compared to spin period and obliquity. Characterization of this anomalous population revealed a strikingly consistent profile: these objects are predominantly larger in size (mean diameter $\sim$64 km vs $\sim$2 km for the non-anomalous majority), reside in remarkably stable, low-inclination ($\sim3^\circ$ vs $\sim9^\circ$), low-eccentricity ($\sim$0.12 vs $\sim$0.16) orbits within the main asteroid belt, and frequently exhibit extreme spin periods that are poorly predicted by the models.

From these results, we learn that the identified anomalous asteroids likely represent a physically distinct population not well-described by evolutionary models based solely on orbit and age, which are implicitly trained on the more numerous population of smaller, more evolved bodies. Their combination of large size and stable orbits strongly suggests they could be surviving primordial planetesimals that have largely escaped significant collisional fragmentation and dynamic excitation since the early Solar System. Their extreme spin states may be a consequence of their size making them less susceptible to the YORP effect, preserving initial spin states or reflecting changes only from rare, large-scale events. This study demonstrates the power of data-driven anomaly detection in identifying objects that lie outside typical evolutionary narratives, providing a prioritized list of targets for future observational campaigns and theoretical investigations aimed at unraveling the mysteries of Solar System formation and evolution.