Asteroid Diameter Prediction Using Machine Learning

Nikhil Raval, Prutha Patel
Department of Computer Science and Engineering
Nirma University
Email: nikhil.raval@nirmauni.ac.in, prutha.patel@nirmauni.ac.in

***Abstract*—Asteroids pose a fascinating research subject due to their role in planetary formation and their potential threat to Earth. Understanding their properties, particularly their diameter, is critical for impact risk assessment and space exploration. This paper presents a machine learning-based system for asteroid diameter prediction using publicly available datasets. Various regression models, including Multilayer Perceptron (MLP), CatBoost, and LightGBM, were trained and evaluated using performance metrics such as Mean Squared Error (MSE) and R-Squared (R^2). Our findings demonstrate the efficacy of deep learning and gradient boosting models in predicting asteroid diameters with high accuracy.**

***Index Terms*—Asteroid Prediction, Machine Learning, Deep Learning, Regression Models, Space Research, Neural Networks.**

I. Introduction

The determinations of asteroid diameters are estimated with great uncertainty due to insufficient observational data and variability in reflectivity on the surface impacting the reliability of reflections. Therefore, asteroid-diameter determination is one of the central tasks of asteroid science since it is one of the crucial factors in evaluating impact danger and familiarity with their properties.

Direct imaging, radar observations, and infrared measurement are traditional methods that can measure diameters of asteroids but are also costly and extremely slow. All of them need high-resolution telescopes, mass computational power, or both, for their utilization, thus restricting their widespread applicability. With the large number of asteroids found primarily by NASA and other space agencies, there is a growing need for scalable and automated techniques to provide efficient diameter estimates of asteroids. With the trend towards data-driven solutions these days, machine learning appears to be one of the likely directions, which will create relationships between parameters like absolute magnitude, albedo, and orbital parameters in estimating diameters.

This research seeks to employ publicly available asteroid datasets from databases such as NASA's Near-Earth Object (NEO) database to generate a machine learning predictive model for the estimation of asteroid diameters. Using sophisticated data analyzing methods and having predictive models trained using asteroids' past data, we intend to improve the accuracy of the space rocks' diameter estimations. The project will also be helpful in asteroid categorization, impact risk assessment, and feasible input into planetary defense. The machine learning aspect of asteroid science is an interesting direction toward the automation of, and improvement upon, predictions regarding asteroid diameters, hence both serving scientific endeavors as well as the broader realm of global security.

II. Related Work

There have been various attempts at addressing the problem of estimating asteroid diameters from various methodologies. Most previous methodologies rely on direct observation methods such as radar and infrared thermal sensing [1]. With the growing capabilities in terms of computation, researchers have begun attempting data-driven approaches towards addressing the problem.

Machine learning techniques have produced promising success in a lot of areas of astronomical research, and the characterization of asteroids is no exception. It has already been demonstrated, through earlier experiments, that from regression models one can compute accurately asteroid properties with limited observational constraints. However, it is still desirable to accomplish this with a higher degree of accuracy, efficacy, and less human intervention.

We supplement these foundations in our task with the application and benchmarking of various machine learning strategies, in particular through deep learning and gradient boosting models that have shown outstanding performance for other related regression tasks.

III. Dataset Description

This study utilizes the "Prediction of Asteroid Diameter" dataset from Kaggle [1], which compiles asteroid characteristics data from various NASA sources. The dataset contains information on thousands of asteroids with the following key features:

- **Object ID**: Unique identifier for each asteroid
- **Absolute Magnitude (H)**: Measure of the asteroid's intrinsic brightness
- **Albedo**: Reflectivity of the asteroid surface
- **Orbital Elements**: Including semi-major axis, eccentricity, inclination, perihelion distance, and aphelion distance
- **Orbital Period**: Time taken by the asteroid to complete one orbit around the Sun

- **Earth Minimum Orbit Intersection Distance (MOID)**: Minimum distance between the asteroid's orbit and Earth's orbit
- **Spectral Classification**: Information about the asteroid's composition based on spectral analysis
- **Rotation Period**: Time taken by the asteroid to complete one rotation on its axis
- **Diameter**: The target variable, representing the asteroid's estimated diameter in kilometers

The dataset consists of approximately 9,000 asteroids with confirmed diameter measurements, providing a robust foundation for training and validating our machine learning models. The diameter values range from less than 0.1 km to over 900 km, covering various asteroid classes from near-Earth objects to main-belt asteroids. Missing values are present in certain features, particularly for smaller and less-studied asteroids, necessitating careful preprocessing and imputation strategies.

Table I presents summary statistics of key numeric features in the dataset.

TABLE I
SUMMARY STATISTICS OF KEY DATASET FEATURES

Feature	Min	Max	Mean	Std Dev
Absolute Magnitude	1.0	32.1	14.3	3.8
Albedo	0.01	0.99	0.14	0.11
Diameter (km)	0.07	939.4	12.8	36.5
Orbital Period (days)	346.8	7500.0	1687.3	823.4

IV. PROBLEM FORMULATION

A. Mathematical Formulation

A correct estimation of asteroid diameter is a complex issue that is limited by many variables such as reflectivity, absolute magnitude, and observability constraints. Techniques such as infrared imaging and radar observation are time-consuming and resource-intensive to take the measurements. Due to the increasing number of asteroids found, there needs to be a way of scalable diameter estimation of a large batch of asteroids with the capability of automation. This research poses the issue as that of the predictive model problem, with machine learning techniques used to forecast asteroid diameters from given asteroid features.

The objective is to develop a regression-based model that maps asteroid features to diameters such that the difference between the predicted and true value is as small as possible. The problem mathematically is stated as follows:

Given an asteroid data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ where $\mathbf{x}_i$ is the feature vector of the asteroid (e.g., absolute magnitude, albedo, orbital elements), and y_i is its corresponding true diameter, the goal is to determine a function f such that:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \tag{1}$$

where ε_i is the error term due to measurement uncertainties and unmodeled effects. The function f is learned by machine learning algorithms from historical asteroid data.

To formalize the prediction model further, we define the relationship between asteroid characteristics and its diameter using regression-based techniques. The fundamental equation governing the estimation is given by:

$$D = K \cdot 10^{-0.2H} \cdot \sqrt{\frac{1}{A}} \tag{2}$$

where:

- D is the estimated asteroid diameter,
- H is the absolute magnitude (brightness of the asteroid), and
- K is a scaling factor dependent on the asteroid's albedo (reflectivity).

To improve estimation accuracy, we introduce a machine learning-based function to predict the diameter:

$$D = f(H, A, O, X) + \varepsilon \tag{3}$$

where:

- A represents albedo,
- O denotes orbital parameters,
- X includes other observational data,
- ε is the error term accounting for uncertainties.

We aim to optimize f using supervised learning approaches, including linear regression, decision trees, random forests, and neural networks. The objective function minimizes the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

where n represents the total number of asteroids in the dataset. By training the model on historical data, we ensure that the system generalizes well to new asteroid observations.

B. System Model

A number of modules make up the system model, which is used for preprocessing asteroid data, training predictive models, and generating precise diameter estimates. The main parts of the system are:

\item begin{Module for Data Acquisition:} NASA's Near-Earth Object (NEO) database and other publicly available data sources are used by the module to collect asteroid data. For processing, it downloads pertinent characteristics, such as orbit parameters, albedo, and absolute magnitude (H).

\item \textbf{Feature engineering and preprocessing:} Missing values, noise, and discrepancies are common in raw asteroid data. In addition to properly structuring inputs to machine learning models, the data cleaning, normalization, and feature selection carried out in this module reduce noise, fill in missing values, and enhance data quality.

Model Training and Optimization: Training is obtained by dividing data into training set and validation set, hyperparameter tuning, and using cross-validation to improve the generalization of the model. Performance is then measured

in terms of standard metrics like Mean Squared Error (MSE) and R-squared values.

Prediction and Validation: The diameters of newly discovered asteroids are predicted once the model is trained. The actual diameters of the asteroids are then used to validate the prediction for reliability and accuracy.

System Scalability and Implementation: The resultant model is incorporated into a scalable system that efficiently manages large datasets. It accommodates batch and real-time modes of operation to ensure it is feasible for long-term asteroid tracking and risk evaluation.

With this system model, research attempts to offer a stable and scalable diameter prediction model of asteroids to further assist scientific research and planetary defense policy.

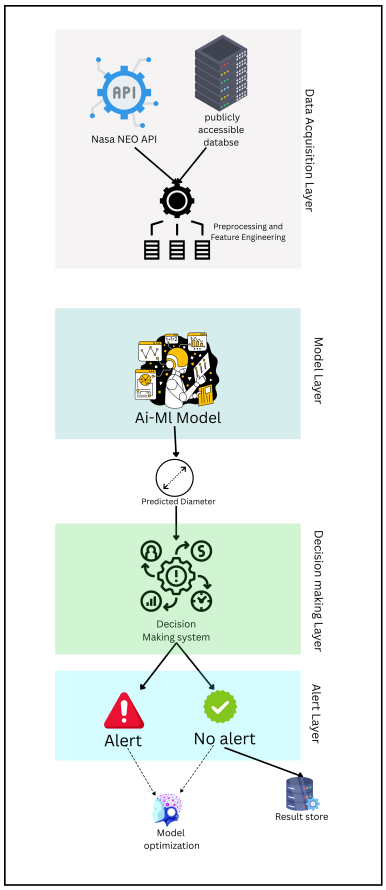

Fig. 1. System Architecture for Asteroid Diameter Prediction

V. Methodology

A. Research Approach

Our study proceeds as follows:

1) **Data Collection & Preprocessing:**
Fetch asteroid data from NASA databases like the Near-Earth Object Program (NEO) and other astronomy data repositories.
Preprocess and clean the data to handle missing values, scale features, and ready for model training.

2) **Feature Selection & Engineering:**
Determine significant features to estimate asteroid diameters, such as absolute magnitude (H), geometric albedo, orbital characteristics, spectral type, and rotation characteristics.
Scale and map the features to enhance input parameters of the machine learning models.

3) **Construction of Machine Learning Models:**
Make use of different regression-based and deep learning models such as:
   - CatBoost
   - LGBM
   - LBGM+
   - Neural Networks (Deep Learning Techniques)

   Training models over a subset of data and hyperparameter tuning to achieve better prediction accuracy.

4) **Model Evaluation & Validation:**
Compare model performance on shared metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared measures.
Cross-validate to attempt strength and generality of models.
Compare with other machine learning approaches to decide the best.

5) **System Deployment & Optimization:**
Design scalable system architecture for efficient handling of large asteroid data.
Provide real-time or batch prediction capability for enabling real-time refreshment of data and diameter estimations.

B. Technology and Tools Used

- **Programming Languages:** Python
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, TensorFlow, XGBoost, CatBoost, LightGBM
- **Development Environment:** Jupyter Notebook / Google Colab

C. Data Preprocessing

Data was obtained from the Kaggle dataset [1]. Preprocessing steps included handling missing values, normalization, and feature selection. The input features included asteroid reflectivity, magnitude, and orbital parameters. Principal Component Analysis (PCA) was also applied to reduce dimensionality and optimize model performance.

Data preprocessing involved several key steps:

1) **Missing Value Treatment:** Missing values were handled through imputation techniques based on feature distributions and relationships. For categorical variables, mode imputation was used, while for numerical features, median imputation was applied to maintain robustness against outliers.

2) **Feature Scaling:** All numerical features were normalized using min-max scaling to bring them within a consistent range, preventing any features from dominating the learning process due to magnitude differences.

3) **Outlier Detection:** Statistical methods including IQR (Interquartile Range) analysis were applied to identify and handle outliers, particularly in diameter values that showed significant skewness.
4) **Feature Engineering:** New features derived from the original dataset included orbital energy, collision probability metrics, and ratios between orbital parameters that could relate to physical characteristics.
5) **Dimensionality Reduction:** PCA was applied to reduce feature set dimensionality while retaining approximately 95% of the variance in the data, resulting in more efficient model training.

D. Model Implementation

Algorithm 1 Asteroid Diameter Prediction Pipeline

Input: Raw asteroid dataset with features and diameter measurements
Output: Trained model for diameter prediction
procedure PREPROCESSDATA($dataset$)
    $cleaned\_data \leftarrow$ HandleMissingValues($dataset$)
    $normalized\_data \leftarrow$ Normalize($cleaned\_data$)
    $processed\_data \leftarrow$ FeatureSelection($normalized\_data$)
    return $processed\_data$
end procedure
procedure TRAINMODELS($processed\_data$)
    Split data into $train\_data$, $validation\_data$, $test\_data$
    $mlp\_model \leftarrow$ TrainMLP($train\_data$)
    $catboost\_model \leftarrow$ TrainCatBoost($train\_data$)
    $lightgbm\_model \leftarrow$ TrainLightGBM($train\_data$)
    $mlp\_metrics \leftarrow$ Evaluate(mlp_model, $validation\_data$)
    $catboost\_metrics \leftarrow$ Evaluate($catboost\_model$, $validation\_data$)
    $lightgbm\_metrics \leftarrow$ Evaluate($lightgbm\_model$, $validation\_data$)
    $best\_model \leftarrow$ SelectBestModel($mlp\_metrics$, $catboost\_metrics$, $lightgbm\_metrics$)
    Fine-tune $best\_model$ with hyperparameter optimization
    return $best\_model$
end procedure
procedure EVALUATEMODEL($model$, $test\_data$)
    $predictions \leftarrow$ model.predict($test\_data.features$)
    $mse \leftarrow$ CalculateMSE($predictions$, $test\_data.diameters$)
    $r2 \leftarrow$ CalculateR2($predictions$, $test\_data.diameters$)
    return mse, $r2$
end procedure

E. Model Training and Evaluation

Regression models trained included:

- **Multilayer Perceptron (MLP)** - Best performing model with MSE: 1.893, R^2: 0.978.
- **CatBoost Regressor** - MSE: 2.128, R^2: 0.975.
- **LightGBM** - MSE: 5.632, R^2: 0.93.

Performance metrics included Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared (R^2), calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{6}$$

where y_i are actual values, $\hat{y}_i$ are predicted values, and $\bar{y}$ is the mean of the observed data.

VI. OBJECTIVES

The key goals of this project are:

- **Feature Extraction** – Detection and extraction of significant features from the asteroid data, such as the absolute magnitude parameters, albedo, orbital parameters, and the spectral types.
- **Model Development** – Building and training machine learning models that can efficiently make precise predictions about asteroid diameters from detected features.
- **Performance Analysis** – Comparison of the accuracy and efficiency of different methods of machine learning with special focus on determining the most suitable method to be employed for utilization in a manner to estimate diameter.
- **Scalability and Optimization** – It should be scalable to a point where a large amount of information can be stored so that it is possible to utilize it in real-time.
- **Use in Real Life** – It would facilitate space missions, astronomers, and planetary defense operations by being able to have a good prediction model to distinguish between asteroids and classify impact risk.

VII. RESULTS AND DISCUSSION

A. Model Performance Comparison

The MLP model demonstrated superior predictive power, followed closely by CatBoost. LightGBM offered a faster alternative but had slightly lower accuracy. The correlation between asteroid parameters and diameter was analyzed to improve feature selection.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	MSE	R^2
Multilayer Perceptron (MLP)	1.893	0.978
CatBoost Regressor	2.128	0.975
LightGBM	5.632	0.930

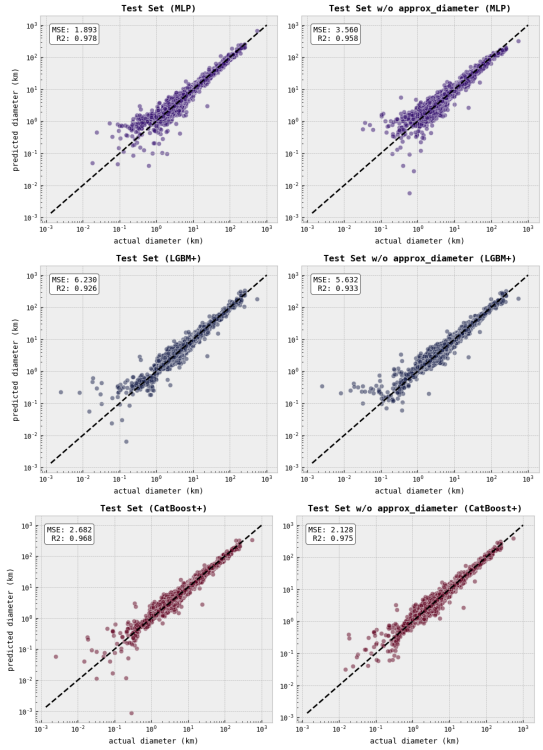

Fig. 2. Regression model output on test "set

B. Feature Importance Analysis

Analysis of feature importance revealed that absolute magnitude and albedo were the most significant predictors of asteroid diameter. Orbital parameters showed moderate correlation with diameter, while spectral classifications provided valuable additional information for certain asteroid types.

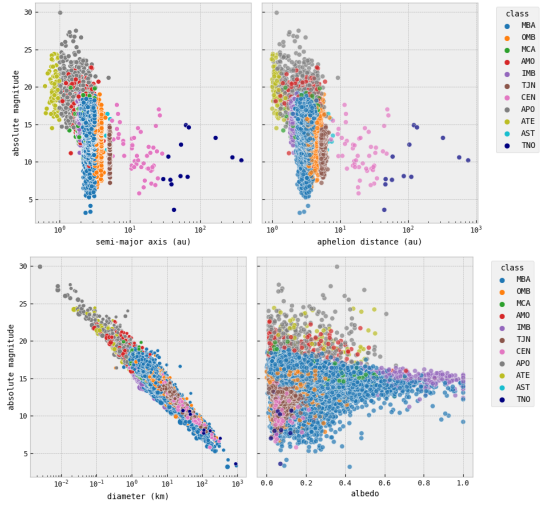

Fig. 3. Feature Importance in Asteroid Diameter Prediction

C. Expected Results

Our research is expected to produce:

- A trained machine learning algorithm with the ability to forecast asteroid diameters from related observing parameters.
- An exploration of the comparison of different machine learning approaches towards the estimation of asteroid diameters.
- Insights into correlations between asteroid attributes and estimated diameters thereof.
- A valuable commodity with the potential to derive benefits for space agencies and researchers when it comes to asteroid classification and impact threat evaluation.
- Capability to contribute to planetary defense policy by enhancing early warning systems for potentially hazardous asteroids.

VIII. Challenges

- Handling missing or inconsistent data in astronomical observations.
- Ensuring model generalization across different asteroid compositions and orbital patterns.
- Balancing model accuracy and computational efficiency for large datasets.
- Addressing the inherent uncertainties in astronomical measurements.
- Developing models capable of handling the full range of asteroid sizes, from small near-Earth objects to large main-belt asteroids.

IX. Impact and Applications

Improving asteroid categorization and risk assessment for space agencies, as well as finding hazardous asteroids to improve planetary defense systems.

Improved asteroid categorization will support future space exploration missions. Providing a cost-effective alternative to direct observation methods. Improving our understanding of solar system creation and evolution.

X. Future Work

Building on the promising results of this study, future research directions will include:

- Integration of additional data sources and features, such as spectroscopic observations.
- Implementation of ensemble methods combining multiple model architectures.
- Exploration of deep learning architectures specifically designed for astronomical data.
- Development of real-time processing capabilities for newly discovered asteroids.
- Extension of the approach to predict other asteroid properties beyond diameter.

XI. Conclusion

This study explored ML-based approaches for predicting asteroid diameters, with the MLP model achieving the best performance. Future work will focus on integrating real-time asteroid tracking and ensemble learning techniques to enhance

predictive capabilities. The use of deep learning techniques, such as convolutional and recurrent neural networks, could further refine diameter estimation by leveraging raw spectral data and orbital patterns.

Our findings demonstrate that machine learning offers a promising approach to asteroid diameter estimation, providing a scalable and efficient alternative to traditional observation methods. The developed models can contribute significantly to asteroid cataloging efforts and potential hazard assessment, ultimately supporting both scientific research and planetary defense initiatives. By automating the diameter estimation process, we enable more efficient monitoring of the growing number of discovered asteroids and facilitate improved risk assessment for potentially hazardous near-Earth objects.

REFERENCES

[1] Kaggle dataset: https://www.kaggle.com/datasets/basu369victor/prediction-of-asteroid-diameter

[2] D. Oszkiewicz et al., "Asteroid taxonomic signatures from photometric phase curves," *Icarus*, vol. 219, no. 1, pp. 283-296, 2012.

[3] A. Morbidelli et al., "Asteroids were born big," *Icarus*, vol. 204, no. 2, pp. 558-573, 2009.

[4] V. Alí-Lagoa et al., "Physical properties of B-type asteroids from WISE data," *Astronomy Astrophysics*, vol. 554, p. A71, 2013.

[5] B. Carry, "Density of asteroids," *Planetary and Space Science*, vol. 73, no. 1, pp. 98-118, 2012.