Team 11 - Near-Earth Object Hazard Detection

Intro to Machine Learning Final Project Report

Alan Mendiola Quezada, Undergraduate Senior in ASE Department
Atharv Vani, Undergraduate Sophomore in SDS Department
Zachary Burdette, Undergraduate Senior in PGE Department

Introduction & Importance

Near-Earth Objects (NEOs) are asteroids or comets with orbits that come within 0.3 astronomical units of the Earth's orbit. [1] While most pose little threat, roughly 9% of these objects are labeled as hazardous due to their potential to cause severe regional or global damage within Earth's atmosphere. Since NEOs travel at high velocities and may be detected shortly before their closest approach, early hazard classification serves as a solution to allow for emergency planning and strategic defensive measures. This study further investigates modern machine learning approaches and questions whether these models can accurately classify whether a newly detected NEO is hazardous using its physical characteristics provided by orbital and observational parameters.

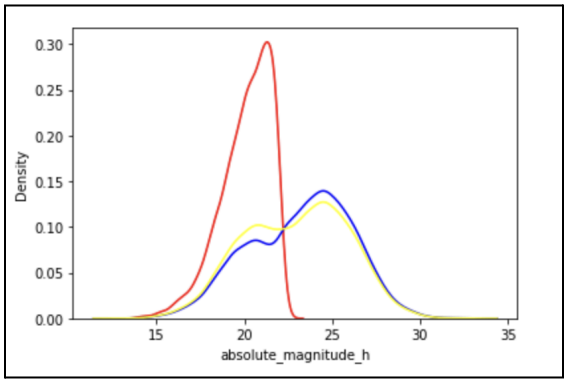
The reason it is important to ask this question is that current detection systems generate large amounts of observational data that require careful analysis. For this study, the team have used data via the NASA Open API gathered by the NASA Center of Near Earth Object Studies (CNEOS) within the Caltech Jet Propulsion Lab. [2] The dataset includes columns such as absolute magnitude, estimated diameter, orbit class type, perihelion distance, aphelion distance, and danger to Earth (which is a binary target). NASA utilizes this data for the main purpose of monitoring objects to make informed decisions. This can be seen in their database as they provide information on objects detected before impact and those with high risk. With so much observational data, current detection systems require reliable classification. Even a small improvement in classifying hazardous objects can benefit early warning systems. However, with the lower proportion of hazardous objects vs non-hazardous, the team will investigate the performance of various machine learning methods and propose a multilayer perceptron (MLP) specialized for the task of classification.

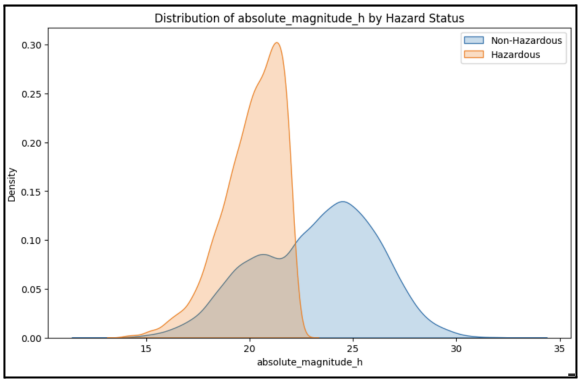
Method & State-of-the-Art (SOTA)

For this data specifically, the existing work was done by Ved Umrajkar, who applied several classical machine learning models to the NASA NEO dataset, yielding the following results. [3] The correlation heatmap for the variables used in the model is provided in **Figure 1**. However, the SOTA pipeline has several limitations that the team aims to improve and address in this study. Rather than filtering out unnecessary identifier information, such as "date" and "description," the team will incorporate feature engineering to model physically meaningful orbital relationships. Additionally, hyperparameter tuning is skipped within the current model, resulting in worse performance. Data leakage issues are also occurring due to improper variables used in the model. The team's approach extends the SOTA models in three key ways as described.

1. *Feature Engineering*

The team engineered new features representing physical characteristics such as mean diameter, eccentricity, orbital radius, observation span, and nonlinear interaction terms(magnitude x diameter). Furthermore, misleading identifier information and time-related data was also removed to improve pattern recognition within physical features. Data preparation and cleaning are essential to this study's design as they allow the model to analyze the important information. It's also important to recognize that the SOTA replaces 15 rows that had "missing/NA" data with the column mean, which can often skew results. Instead of replacing with the mean, those entries were dropped, causing a small distribution change, contributing to the improvements of the model. The columns lacking data are "diameter max," "diameter min," "absolute magnitude," "last observation date," and "first observation date." This outcome is also supported by the comparison of variable distribution graphs. The left plot shows the SOTA, and the right plot shows the changes after dropping the entries stated above.





2. *Hyperparameter Tuning*

Using random search and grid search, the team optimized the four models, resulting in improved predictive performance for all models except for the logistic regression model.

Random search is an efficient hyperparameter tuning technique that samples random combinations from predefined parameter distributions instead of testing every option. This approach is effective when the hyperparameter space is large or when only a few parameters significantly influence model performance. By exploring a wide and diverse set of configurations with fewer evaluations, random search often discovers high-performing models more quickly than grid search. In this study, models such as the Decision Tree, Random Forest, and Logistic Regression benefited from random search because they contain multiple continuous or wide-ranged hyperparameters provided below.

Grid search is another hyperparameter tuning method that evaluates every possible combination within a fixed set of parameter values. Since it exhaustively tests all configurations, grid search provides stable, interpretable results and ensures that no potential combination within the grid is missed. For this project, the team used grid search for Linear Discriminant Analysis (LDA) since it has relatively few hyperparameters, shrinkage and solver, making exhaustive search computationally practical and reliable.

Decision Tree	Random Forest	Logistic Regression	LDA
Depth: 7	Depth: 9	C: 7.32	Shrinkage: 0.1
Samples Leafs: 9	Sample Leafs: 2	Penalty: L2	
Sample Split: 9	Sample Split: 4	Solver: Saga	Solver: lsqr
	Features: 0.5		

3. Neural Network - MLP

The team also included a neural network model by training a multilayer perceptron. Though MLPs have been used for various tasks, the team references models used to solve other problems and designed a neural network for the problem statement. This is an existing method that hasn't yet been explored in the context of NEO classification. The team proposed and developed a neural network, as the part lacking in the SOTA models is the ability to capture nonlinear relationships more effectively. A report of the neural network trained on the data has been provided in **Figure 2**.

Contributions

The team's contributions to the existing body of literature come from the unique approach when handling features and hyperparameters. Unlike prior SOTA work, which used raw features with minimal preprocessing, the team developed a meaningful engineered feature set that captures orbital relationships, nonlinear interactions, and observational properties more effectively. The results of the model's accuracy improvement are described in the table shown in the following section.

One focus of this study is the contributions presented by using a neural network. By adding an MLP model, a baseline is provided for more deep learning approaches, such as LSTM or physics-informed models. Most public NEO hazard classifiers rely on decision trees or logistic regression. The neural network is unique as it adds nonlinear modeling, multi-layer representation learning, and scaled, normalized astrophysical feature inputs. This shows how deep learning can be applied to NEO risk prediction.

The MLP architecture consisted of two hidden layers with 64 and 32 neurons using ReLU as an activation function. It's also important to consider limitations because neural networks are sensitive to feature scale. To account for this, the team standardized all continuous variables before training to achieve ideal convergence. The design allows the network to learn complex interactions between orbital parameters. These are relationships that decision trees and linear models cannot represent as effectively. Although the MLP required more computation and careful tuning, it achieved a competitive accuracy of 92.1%, outperforming most classical models and validating the usefulness of deep learning for astrophysical classification problems.

Results/Findings

Model	State of the Art(SOTA)	Hyperparameter + Feature Engineering
Decision Tree	90.7%	91.7%
Random Forest	91.9%	92.5%
Logistic Regression	91.8%	91.7%
LDA	90.5%	91.4%
MLP(64, 32)	-	92.1%

From the accuracy scores measured on holdout sets, Random Forest is proven to be the best classical model and improves an additional 0.6% after feature engineering and tuning to yield a 92.5% correct classification rate. Meaningful engineered features such as the “mean diameter,” “orbit radius,” and “eccentricity” are important predictors in the model, confirming the hypothesis that physical relationships between orbital parameters improve hazard classification.

It can also be observed that the neural network offers an advantage as it captures nonlinear features automatically and performs well as the feature space grows. However, it does require more computing power and training time. This suggests room for work in the future regarding deep learning approaches by carefully managing parameters to improve performance. A detailed analysis/report of the neural network is provided in **Figure 2**.

Random Forest remained the top-performing model due to its ability to capture nonlinear boundaries and feature interactions without heavy preprocessing. In comparison, Logistic Regression slightly declined after tuning, likely because the engineered features introduced nonlinear relationships that violate logistic regression’s linear separability assumptions. Understanding how these models work allows the team to interpret meaningful scientific results from them. Feature importance rankings confirm that orbital geometry (eccentricity, perihelion distance, orbit radius) and size-related metrics (mean diameter) are the strongest indicators of hazard potential. The MLP performed competitively, suggesting that deeper neural approaches may eventually surpass ensembles once larger datasets or augmented hazardous samples become available.

Conclusion/Outlook

In this study, the team analyzed and developed an improved machine learning pipeline for classifying hazardous NEOs. By refining models introduced by the SOTA with engineered features and careful tuning, classification accuracy was increased. Variables such as “mag_x_diameter” capture nonlinear features, helping the model recognize interaction between size and orbital eccentricity and variability in orbital radius’s effect on hazard likelihood. Furthermore, the proposed MLP is the second-best model for this task, suggesting the exploration of more deep learning approaches. The team’s work demonstrates that preprocessing data can significantly enhance early detection capabilities and drive strategic efforts for safety.

Future research should be done to improve the accuracy of these models for the classification task. One possibility is expanding the dataset with more hazardous objects to combat class imbalance. More observational data could be useful in helping the model recognize hazardous objects. Although it's impossible to "create observations", drawing on different data sources for physical characteristics of NEOs is a possibility, as NASA's API has a wide range of functionality. The performance gap between the Random Forest and the MLP highlights the value of nonlinear modeling while also revealing that deep learning models require bigger and balanced datasets to reach full potential. The results indicate that combining sequence-aware neural models with richer orbital time-series data to capture NEO trajectory evolution directly is an optimal route to explore. Overall, this study presents results to drive strategic efforts by improving classification models used in a real-world system for NEO detection and hazard/risk mitigation.

Appendix

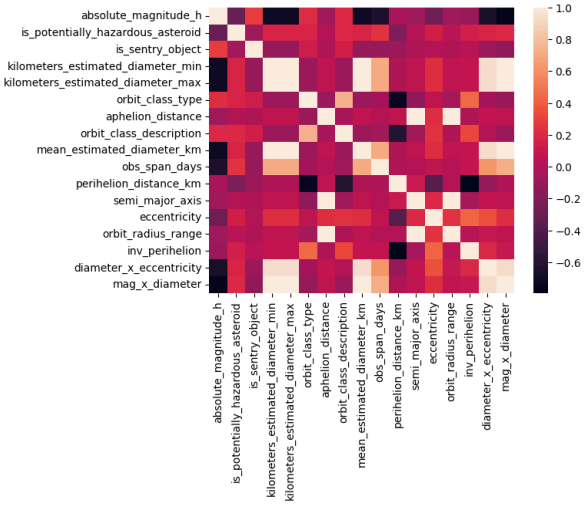


Figure 1. Physical Characteristics Correlations

```
Neural Network Accuracy: 0.9212007504690432

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.97      0.96      4378
           1       0.57      0.38      0.46       419

    accuracy                           0.92      4797
   macro avg       0.76      0.68      0.71      4797
weighted avg       0.91      0.92      0.91      4797


Confusion Matrix:
[[4261  117]
 [ 261  158]]
```

Figure 2. MLP Classification Report

References

[1] Robert.wickramatunga, “United NationsOffice for Outer Space Affairs,” Near-Earth Objects, https://www.unoosa.org/oosa/sk/ourwork/topics/neos/index.html (accessed Dec. 1, 2025).

[2] “Center for Neo Studies,” NASA, https://cneos.jpl.nasa.gov/ (accessed Dec. 1, 2025).

[3] A. Ramachandran, “NASA Near Earth Objects Information,” Kaggle, https://www.kaggle.com/datasets/adityaramachandran27/nasa-near-earth-objects-information/data (accessed Dec. 1, 2025).