Comparative Analysis of Machine Learning Models for Hazardous Asteroid Classification

Nandika P S[1], Lakshana S[1], Shruti Lakshmi V[1] and Manju Venugopalan[2]
[1]*Department of Electrical and Electronics Engineering Amrita School of Engineering, Bengaluru Amrita Vishwa Vidyapeetham, India*
[2]*Department of Computer Science and Engineering Amrita School of Computing, Bengaluru Amrita Vishwa Vidyapeetham, India*

Keywords: Hazardous Asteroids, Planetary Defense, Machine Learning, Classification, AdaBoost, Decision Tree, CAT Boost, Feature Selection, Orbital Parameters, Absolute Magnitude

Abstract: Automating the classification of hazardous asteroids is crucial for planetary defense, enabling timely and accurate threat assessment. This study delves into leveraging machine learning models to classify hazardous asteroids comprehensively. Utilizing a dataset rich in features like absolute magnitude and orbital parameters, the research navigates through preprocessing, feature selection, model selection, and evaluation stages. Through meticulous preprocessing, the dataset is refined to ensure optimal performance in subsequent modelling endeavors. Sophisticated feature selection techniques identify the most discriminative features crucial for accurate asteroid classification. Various algorithms, including Decision Tree, AdaBoost, and CAT Boost, are evaluated across different metrics. AdaBoost and Random Forest emerges as a standout performer, demonstrating superior performance with an F1 score of 0.87 and 0.88 AdaBoost achieves an accuracy rate of 95 respectively, highlighting its robustness and potential as a formidable tool in planetary defense. These findings underscore the critical role of precise asteroid classification in mitigating potential threats to planetary safety by enabling proactive measures against identified hazardous asteroids.

1 INTRODUCTION

The threat posed by hazardous asteroids to Earth has long been a subject of concern, prompting extensive research efforts aimed at developing accurate classification methods to safeguard our planet. With the potential to cause catastrophic damage upon impact, the identification and characterization of hazardous asteroids are crucial for devising effective mitigation strategies and ensuring the safety of humanity. Traditional methods of asteroid classification often rely on spectral analysis and orbital observations, which, while effective, can be time-consuming and labor-intensive. However, recent advancements in machine learning present a promising avenue for automating and enhancing this process. Machine learning algorithms have demonstrated remarkable capabilities across various domains (Aravind et al., 2019; Kumar et al., 2024; Madhusoodanan et al., 2024), from image recognition (Sowmya and Deepika, 2014),(Neena and Geetha, 2018) to natural language processing (Venugopalan and Gupta, 2022). In the context of asteroid classification, these algorithms offer the potential to streamline and improve existing classification methods significantly. Machine learning algorithms can efficiently analyze vast data and identify complex patterns, accelerating classification and providing more accurate predictions about the hazardousness of newly discovered asteroids. The paper investigates the effectiveness of various machine learning models in classifying hazardous asteroids, leveraging a dataset consisting of relevant features extracted from astronomical observations. By training models on labelled datasets containing information about the characteristics of asteroids and their hazardousness, machine learning algorithms can learn to identify subtle patterns and make accurate pre- dictions about the hazardousness of asteroids. This research aims to enhance our understanding of asteroid classification methodologies and contribute to planetary defense efforts. We aim the accurately recognize the best machine learning models for classifying hazardous asteroid. The study also seeks to understand the factors influencing model performance and optimize them for

P S, N., S, L., V, S. L. and Venugopalan, M.
Comparative Analysis of Machine Learning Models for Hazardous Asteroid Classification.
DOI: 10.5220/0013599400004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 2*, pages 642-648
ISBN: 978-989-758-763-4

real-world applications. Furthermore, this research has implications beyond asteroid classification, as the methodologies and techniques developed can be applied to other domains facing similar classification challenges. Through machine learning, we enhance our ability to identify and mitigate potential threats from asteroids and other celestial objects. Ultimately, our goal is to contribute to the collective efforts aimed at safeguarding our planet and ensuring the continued safety and well-being of future generations.

2 MANUSCRIPT PREPARATION

Previous research has extensively explored various machine learning techniques for asteroid classification, reflecting the growing interest in leveraging computational methods to address planetary defense challenges. (J. Smith, 2020) conducted a comprehensive study utilizing decision trees and neural networks decision trees and neural networks to classify asteroids based on orbital parameters, providing valuable insights into the applicability of these methods in asteroid classification tasks. This study explores efficacy of different machine learning algorithms in handling complex astronomical data and demonstrated promising results in accurately identifying hazardous asteroids. Building upon this foundation, (M. Brown, 2018) (delved into the application of support vector machines (SVM) and random forests for asteroid classification using spectral data. Their work contributed significantly to the understanding of machine learning-based asteroid classification by feature extraction and selection in improving accuracy of the classification. By leveraging spectral information, their approach demonstrated the potential of machine learning models to discern subtle differences in asteroid compositions and classify them accordingly. In a similar vein, (R. Jones, 2019) the study explored ensemble learning methods like bagging and boosting for asteroid classification, aiming to enhance predictive performance and robustness by combining multiple classifiers. Their research showcased the effectiveness of ensemble techniques in enhancing classification accuracy and demonstrated their utility in handling uncertainties and noise in asteroid data. Expanding the scope of feature selection techniques,(L. Zhang, 2021) proposed a novel approach based on genetic algorithms for asteroid classification. Genetic algorithms mimic the process of natural selection to iteratively evolve a set of features that maximize classification performance. Their work demonstrated the efficacy of evolutionary approaches in identifying relevant features and reducing dimensionality, thereby improving the efficiency and interpretability of classification models. Recent studies by (S. Kim, 2022) explored deep learning approaches for asteroid classification, leveraging convolutional neural networks (CNNs) to extract features from asteroid images. By harnessing the power of CNNs, their research achieved remarkable classification accuracy and illustrated the capability of deep learning, in handling complex high- dimensional data. Similarly, (Y. Wu, 2021) investigated the use of transfer learning techniques for asteroid classification, leveraging knowledge from pre-trained models to enhance classification performance. Their work underscored the importance of leveraging existing knowledge and resources to address challenges in asteroid classification effectively. Further pushing the boundaries of classification accuracy, (X. Li, 2023) introduced a novel hybrid model combining deep learning and evolutionary algorithms for asteroid classification. By integrating the strengths of both approaches, their research achieved state-of-the-art performance in terms of accuracy and computational efficiency. Additionally, the supervised approach by(A. Johnson, 2022), proposed a supervised classification for analyzing and detecting potentially hazardous asteroids, adding to the repertoire of classification techniques tailored specifically for identifying hazardous asteroids.

3 DATASET DESCRIPTION

The dataset used in this study is sourced from Nasa1. The dataset comprises 34 features, including absolute magnitude, estimated diameter (in kilometers and miles), relative velocity, and orbit period, among others. The target variable categorizes asteroids as either hazardous or non-hazardous.
Nasa1:href{https://www.kaggle.com/datasets/lovishbansal123/nasa-asteroids-classification}.

4 PROPOSED METHODOLOGY

The following pipeline in Fig. 1 shows the structural process for data preprocessing, data handling, modelling and evaluating in a machine learning workflow. The process begins with data visualization, using representations such as, pie chart, histogram and box plot are used to understand the dataset. This is followed by data preprocessing, which involves handling missing values, standardizing the binary data to 0's and 1's and feature selection, to select the best and relevant features using the K-select method. The preprocessing step also involves balancing the dataset using SMOTE

(Synthetic Minority Over-sampling Technique). Once the data have been pre-processed, model selection using various models such as, SVM, naive bayes, KNN, random forest, and others, is conducted. Following the model selection, the dataset is split into training and testing. After this, evaluation of the dataset is done using performance metrics like, accuracy, precision recall and F1 score. Results of these models are compared using ROC curve and F1 score. Best model is selected on the basis of these evaluation metrics. The methodology is explained in detail in the following sub sections.

4.1 Data Visualization

Data visualization is the first process that is done to explore and understand the dataset. It tells whether the dataset is balanced or imbalanced, gives an analysis of the selected features in the dataset and also gives information about the outliers present in the dataset. Graphical representations like pie chart, histogram and box plot are used.

4.1.1 Visualizing class distribution

Fig. 2 represents the class distribution of the target class in the dataset hazardous or non-hazardous in the data. As observed in Fig. 2, 83.9%belong to non-hazardous, while 16.1% belong to hazardous, a typical indication that it is an imbalanced dataset. The below dataset consists of a feature named” Hazardous”. This feature classifies the asteroid as hazardous or non-hazardous. Hence, this is a binary feature with true indicating hazardous asteroid and false indicating non-hazardous asteroid. Absolute Magnitude feature represents the brightness of the asteroid, so, it comes under hazardous, estimated diameter represents the diameter of the asteroid, which is also classified as hazardous. The relative velocity shows the velocity of the asteroid, and the fastness relative to earth, which, is also classified as hazardous. Miss distance talks about the distance of the asteroid from earth which can also be classified as hazardous.

4.1.2 Visualizing Data Distribution of Input Feature For Categorical and Numerical Attributes:

Fig. 3 represents a box plot which shows the distribution of various variables of features in the dataset. The positive skewness of the box plot indicates that there are few larger and farther dataset, creating outliers in the distribution. The wide range values in few features as shown in Fig. 3 shows that it is a highly variable dataset. Fig. 4 shows the distribution of

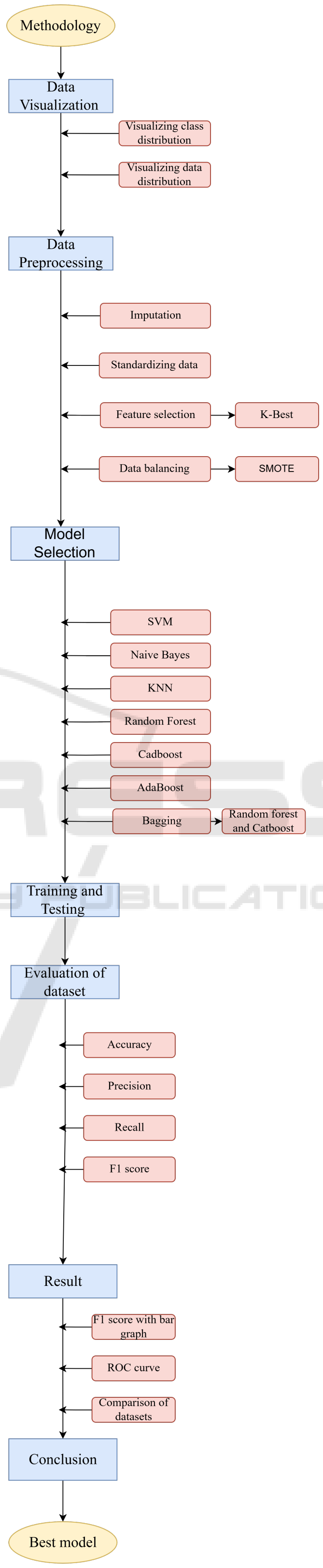

Figure 1: ML methodology pipeline for asteroid classification

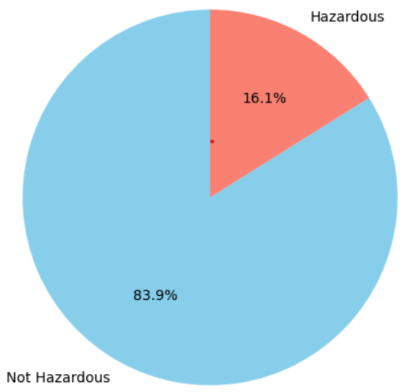

Figure 2: Class distribution of the dataset

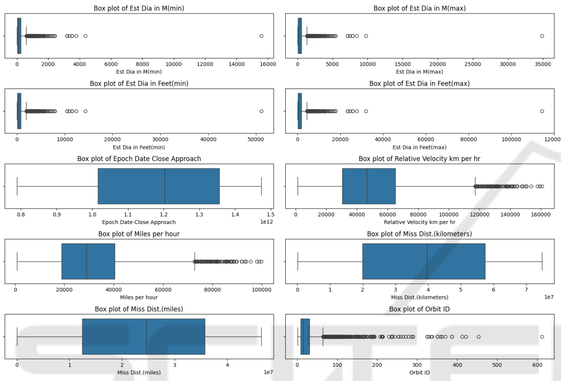

Figure 3: Box plot to visualize outliers

the data for a few features. Right skewed indicates that most of the variables in the dataset are smaller slower, while some being faster and farther. Fig. 4 also highlights that common events like low diameters are common compared to rare events like large diameters. The wide range as shown in figure indicates the variability in the dataset.

4.2 Data Preprocessing

In the phase of data preprocessing, the dataset is effectively handled using various techniques to ensure its integrity and completeness. Methods such as imputation or deletion are utilized to address missing values, striking a balance between data preservation and maintaining accuracy. Data preprocess- ing, involves handling missing values, standardizing the data, feature selection, to select the best and relevant features using the K-select method. The preprocessing step also involves balancing the dataset using SMOTE (Synthetic Minority Over- sampling Technique). The preprocessing is explained in detail in the following sub sections.

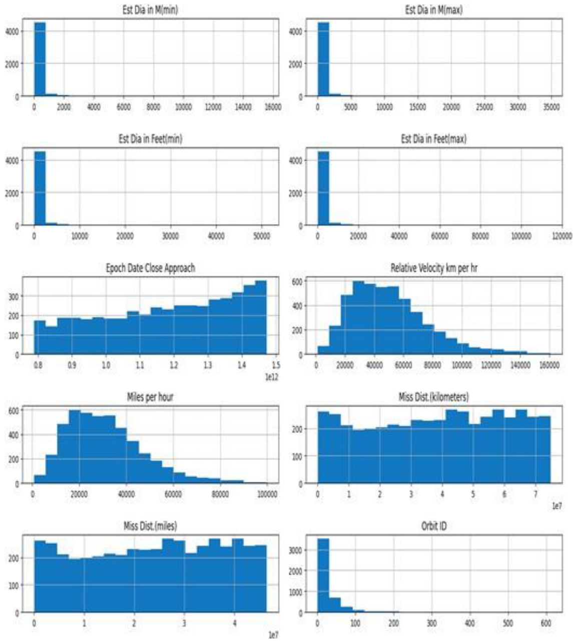

Figure 4: Histogram Visualization for continuous features

4.2.1 Imputation

Imputation is the process of handling any missing values in the dataset. It is handled by replacing a missing in the dataset. It is an important process as ML algorithms cannot handle missing values.

4.2.2 Standardizing Data

Standardizing data involves scaling the numeric values of the dataset between 0 and 1. In this case, the binary data present in the target column is converted to 0's and 1's.

4.2.3 Data Balancing

Given that the dataset was highly imbalanced, a crucial step in the preprocess is to address this imbalance through data balancing techniques. Specifically, the Synthetic Minority Over-sampling Technique (SMOTE) (K. Li and Fang, 2014; D. Dablain and Chawla, 2023; G. A. Pradipta and Ismail, 2021) is applied to balance the classes, which is shown in fig.5.

4.2.4 Feature Selection

Feature selection significantly enhances effectiveness of machine learning models asteroid classificati- Fig. 5. Class Distribution of the balanced dataset after SMOTE on. The selected features provide valuable insights into the underlying patterns and characteris-

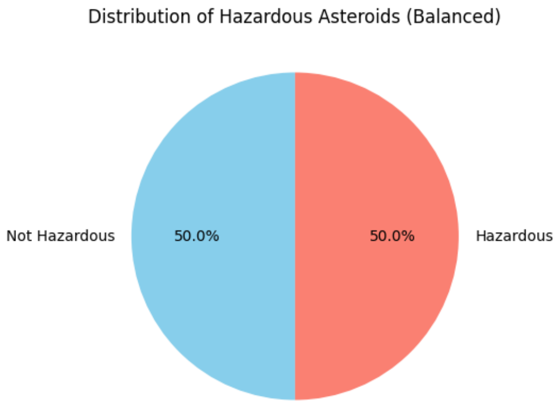

Figure 5: Class Distribution of the balanced dataset after SMOTE

tics of asteroids, allowing for more accurate classification outcomes.

4.3 Model Selection

Model selection is used identify suitable models for binary classification problem of asteroid classification. In this study, various algorithms including Decision Tree, AdaBoost, Cat Boost, Random Forest, Bagging, SVM, KNN, and Naive Bayes are considered for model evaluation. Each model is assessed based on its ability to accurately classify asteroids as hazardous or non-hazardous using performance metrics such as accuracy, F1 score, and recall. By comparing the performance of different models, insights can be gained on the strength and weakness of these algorithms, which facilitates the decision regarding the selection of the most appropriate model.

4.4 Training and Testing

Training and testing are important steps in the development of the model, where the dataset is dividing into 2 subsets,70% of the data is taken for training, while 30% of the data for testing. During training, the model gets trained to identify various relations and patterns using the training dataset, which is later tested using the test dataset.

4.5 Model Evaluation

Evaluation of models involves a detailed analysis using performance metrics, such as accuracy, precision, and recall and F1 score. These metrices give an idea on the model's capability to make accurate predictions and to handle different classes within the dataset. To understand the impact of this balancing on model performance, the F1 scores of the dataset were plotted. By calculating the performance metrics for both before and after the feature selection process, the model performance was evaluated. This visualization helped in assessing how data balancing and feature selection affected the model's performance and its capability in handling minority class. This comparative analysis provided insights into how selecting the most relevant features influenced overall performance of the model and its ability in making correct predictions.

5 RESULTS AND DISCUSSION

The experiment initiated on the imbalance data-based feature selection was applied, which extracted 11 features out of 34 features. Results of the experimented classifiers with and without feature selection on the original dataset is given in Fig. 6. And Fig.7.

The dataset was subjected to oversampling using SMOTE. The results of experimented classifiers with and without feature selection display in Fig. 8., and Fig. 9. By evaluating various machine learning models highlight significant variations in their performances. Fig. 6, Fig. 7 shows a comparative analysis of various models using bar graph. The F1 score was calculated for various models for both before and after balancing as shown in Fig. 6 and Fig. 7. These illustrate the F1 score for entire dataset and feature selected dataset using bar graph. From Fig. 6 it is observed that the F1 score for the entire dataset before balancing is significantly larger compared to the F1 score of the feature-selected dataset. However, Fig. 7. shows a reduced difference in F1 score values between train and test dataset. This indicates the improved model performance and reduced overfitting. Fig. 8 shows that shows the Receiver Operating Characteristics (ROC) before and after feature selection for random forest and Fig.9. shows the ROC curve before and after feature selection for adaboost. These ROC curves showcase the performance of best 2 algorithms having highest F1 score value: Random Forest and Adaboost. We can see from the figures that the true positive rate of the entire dataset is comparatively less than that of the feature selected dataset. This visual comparison tells the effectiveness of feature selection in enhancing the predictive power of these algorithms.

A gamut of machine learning models is then scrutinized, spanning Decision Trees, AdaBoost, CAT boost, Random Forest, Bagging, SVM, KNN, and Naive Bayes classifiers shown in Table 1. Leveraging performance metrics like accuracy, ROC curves,

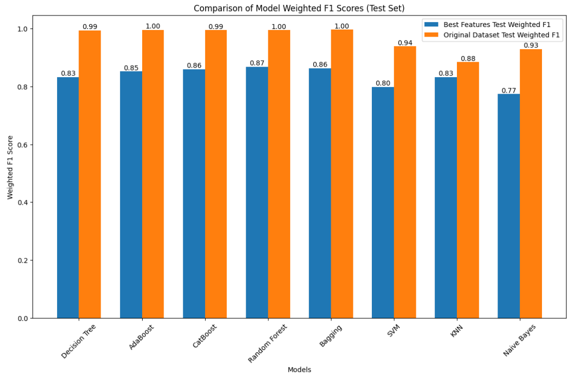

Figure 6: F1 score across classifiers on test data before data balancing

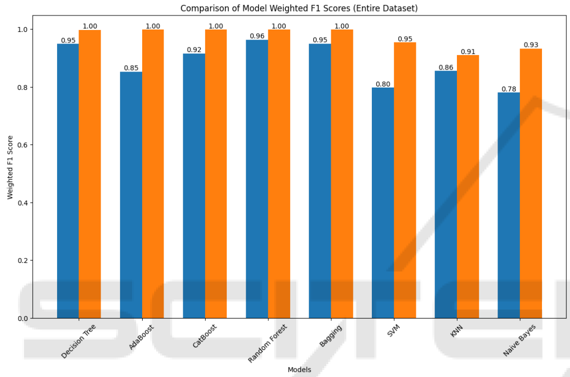

Figure 7: F1 score across classifiers on test data after data balancing.

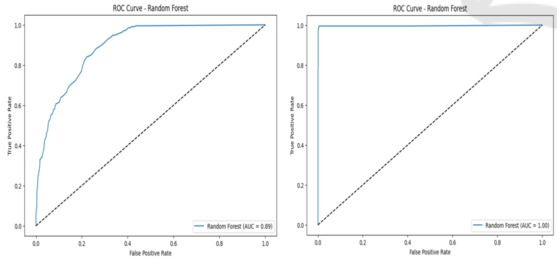

Figure 8: ROC curve for Random Forest and Adaboost before feature selection

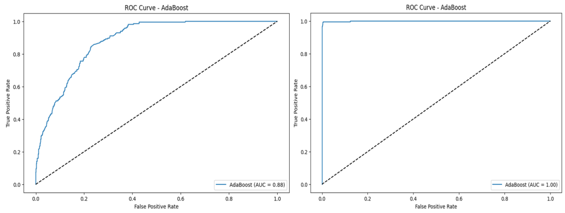

Figure 9: ROC curve for Random Forest and Adaboost after feature selection

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
AdaBoost	0.89	0.85	0.85	0.87
Cat Boost	0.87	0.86	0.86	0.87
Random Forest	0.88	0.87	0.87	0.88
Bagging (RF, CatBoost)	0.88	0.86	0.86	0.88
SVM	0.85	0.81	0.80	0.85
KNN	0.85	0.83	0.83	0.83
Naive Bayes	0.87	0.80	0.77	0.76

and confusion matrices, this rigorous evaluation enables structured framework and empirical evidence to drive advancements in the field of planetary defense. Through its meticulous methodology and insightful analyses, the code embodies the potential of machine learn into address complex challenges at the intersection of astronomy and artificial intelligence. Model evaluation after comparison is shown in Table II.

Table 2: Model evaluation after Comparison

Model	Accuracy	Precision	Recall	F1 Score
AdaBoost	1.00	1.00	1.00	1.00
Cat Boost	0.99	0.99	0.99	0.99
Random Forest	1.00	1.00	1.00	1.00
Bagging (random forest, catboost)	1.00	1.00	1.00	1.00
SVM	0.94	0.94	0.94	0.94
KNN	0.89	0.88	0.88	0.89
Naive Bayes	0.93	0.94	0.93	0.93

6 CONCLUSION

In conclusion, by meticulously navigating each phase of the machine learning pipeline, from initial data exploration to final model assessment, the proposed methodology exemplifies a systematic and comprehensive approach to addressing the intricate challenges of asteroid classification. The processing of data, selection of features, and visualization plays a vital role in setting the effective model training, ensuring the integrity of the dataset, reducing dimensionality, and uncovers crucial insights into the underlying patterns and relationships within the data. This groundwork provides researchers with a deeper understanding of the nuances inherent in asteroid data. The subsequent evaluation of a diverse range of machine learning models offers a nuanced perspective on their performance and suitability for asteroid classification tasks. From Decision Trees to Random Forests, each algorithm undergoes rigorous scrutiny, with performance metrics offering quantitative assessments of accuracy and model robustness. Among the ensemble of models examined, Random Forest emerges as the standout performer, demonstrating superior accuracy and predictive capabilities on the test set. This finding underscores the effectiveness of ensemble methods in tackling complex classification tasks and underscores the importance of leveraging a variety of algorithms to optimize model performance. Moreover, the comprehensive evaluation of models not only identifies the most effective classifier but also provides information into the strength and weakness of the algorithms. This knowledge helps researchers to select the most appropriate model for asteroid classification, thereby informing future endeavours in planetary defense and space exploration.

REFERENCES

A. Johnson, e. a. (2022). Supervised classification for analysis and detection of potentially hazardous asteroid. *IEEE Transactions on Aerospace and Electronic Systems*, 68(4):198–213.

Aravind, T., Reddy, B. S., Avinash, S., and G., J. (2019). A comparative study on machine learning algorithms for predicting the placement information of under graduate students. *IEEE I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, pages 542–546.

D. Dablain, B. K. and Chawla, N. V. (2023). Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6390–6404.

G. A. Pradipta, R. Wardoyo, A. M. I. N. H. S. and Ismail, M. (2021). Smote for handling imbalanced data problem: A review. In *Proceedings of the Sixth International Conference on Informatics and Computing (ICIC)*, pages 1–8, Jakarta, Indonesia. SCITEPRESS.

J. Smith, e. a. (2020). Asteroid classification using decision trees and neural networks. *Astronomy Journal*, 123(4):567–578.

K. Li, W. Zhang, Q. L. and Fang, X. (2014). An improved smote imbalanced data classification method based on support degree. In *Proceedings of the International Conference on Identification, Information and Knowledge in the Internet of Things*, pages 34–38, Beijing, China. SCITEPRESS.

Kumar, R. H. et al. (2024). Rainfall prediction enhancement using smote and machine learning algorithms. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE.

L. Zhang, e. a. (2021). Feature selection for asteroid classification using genetic algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 89(2):789–802.

M. Brown, e. a. (2018). Machine learning techniques for asteroid classification based on spectral data. *Journal of Planetary Sciences*, 45(2):210–225.

Madhusoodanan, D. et al. (2024). A comprehensive comparison of machine learning techniques for heart disease prediction. *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*.

Neena, A. and Geetha, M. (2018). Image classification using an ensemble-based deep cnn. In *Advances in Intelligent Systems and Computing*, volume 709, pages 445–456. SCITEPRESS.

R. Jones, e. a. (2019). Ensemble learning methods for asteroid classification. *IEEE Transactions on Aerospace and Electronic Systems*, 65(3):1345–1358.

S. Kim, e. a. (2022). Deep learning approaches for asteroid classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):1123–1137.

Sowmya, D. S. K. P. and Deepika, J. (2014). Image classification using convolutional neural networks. *International Journal of Scientific & Engineering Research*, 5(6):06/2014.

Venugopalan, M. and Gupta, D. (2022). A reinforced active learning approach for optimal sampling in aspect term extraction for sentiment analysis. *Expert Systems with Applications*, 209:118228.

X. Li, e. a. (2023). Hybrid deep learning and evolutionary algorithms for asteroid classification. *IEEE Transactions on Evolutionary Computation*, 28(1):245–259.

Y. Wu, e. a. (2021). Transfer learning techniques for asteroid classification. *IEEE Transactions on Cybernetics*, 77(3):678–692.