Veermata Jijabai Technological Institute, Mumbai

Coursework Project on **Exoplanet Classification System**

Department of Electrical Engineering
Veermata Jijabai Technological Institute
Mumbai, Maharashtra, India

Electronics Engineering

Veermata Jijabai Technological Institute, Mumbai

Certificate of Submission

Certified on ____________that the project titled "**Exoplanet Classification System**" has been submitted by the following project group of Final year students of Electronics Engineering:

1 Shashwat Barai (221060007)

2 Sachin Vishvakarma (211060030)

Signature: _______________

Guide: Prof Amey Nandgaonkar

Affiliation: Department of Electrical Engineering.
Veermata Jijabai Technological Institute, Mumbai, Maharashtra, India.

Abstract

The discovery of exoplanets—planets orbiting stars outside our solar system—is a cornerstone of modern astrophysics. However, the primary method of detection, Transit Photometry, is hindered by the vast volume of noise in astronomical data and the extreme rarity of planetary events. Traditional manual inspection and statistical thresholding methods are becoming increasingly inefficient as data volumes from missions like Kepler and TESS grow exponentially.

This project proposes an automated Exoplanet Classification System using Neural Networks. By leveraging a Deep Artificial Neural Network (ANN) and digital signal processing techniques, specifically the Fast Fourier Transform (FFT), the system transforms raw light intensity data from the time domain to the frequency domain. This transformation effectively isolates periodic planetary signals from stochastic stellar noise. To address the severe class imbalance (0.7% positive class), the project implements SMOTE (Synthetic Minority Over-sampling Technique) and class-weighted optimization. The final model achieves a ROC AUC score of >0.99 and 100% Recall on unseen test data, demonstrating its potential as a robust, scalable filtering tool for next-generation astronomical surveys.

Table of Contents

Introduction

1.1 Problem Statement

The search for extraterrestrial life and habitable worlds is one of the most profound scientific endeavors of the 21st century. While traditional astronomical methods for detecting exoplanets rely on manual analysis of light curves or standard statistical thresholding, these methods are becoming increasingly bottlenecked by the sheer volume of data. Modern space telescopes like Kepler, TESS, and the James Webb Space Telescope generate massive amounts of time-series data representing the light intensity (flux) of stars.

The core challenge lies in the "Transit Method" of detection. When a planet passes in front of its host star, it causes a minute dip in brightness. Identifying these dips is notoriously difficult due to three primary factors:

1. Cosmic Noise: Interference from background radiation, thermal noise, and instrument error can mask faint signals.
2. Stellar Variability: Natural fluctuations in a star's brightness (e.g., sunspots, flares, or pulsations) can mimic or obscure planetary transits.
3. Rarity of Events: The vast majority of observed stars do not host planets (or at least not detectable ones), creating an extreme class imbalance that biases standard machine learning algorithms.

1.2 Proposed Solution

This project proposes an automated, scalable approach leveraging Deep Neural Networks (ANNs). By training models on large datasets of star flux readings, the system learns to distinguish the underlying periodic patterns of exoplanet transits from stochastic noise.

The solution integrates digital signal processing via Fourier Transforms to shift analysis from the time domain to the frequency domain, making periodic signals significantly easier to detect. This results in a system capable of filtering and classifying potential exoplanet candidates with high precision, significantly reducing the manual workload for astronomers .

Theoretical Framework

2.1 Light Flux Intensity Analysis

- Definition: Light Flux Intensity is the fundamental unit of measurement in the dataset. It represents the varying brightness collected from a star over discrete time steps. In the Kepler dataset, columns FLUX.1 through FLUX.3197 represent consecutive brightness readings.
- Behavioral Patterns:
    - Non-Exoplanet Stars: These typically exhibit stochastic noise or "flat-line" behavior. Any fluctuations are generally random (Gaussian noise) or follow long-term stellar cycles that are unrelated to planetary orbits.
    - Exoplanet Stars: These stars show specific, structured anomalies known as "dips." However, because planets are extremely small compared to their host stars, these dips are often microscopic (e.g., less than 0.01% drop in brightness), making them incredibly hard to detect in the raw time domain against the backdrop of high-variance noise.

2.2 Transit Photometry & Periodic Dips

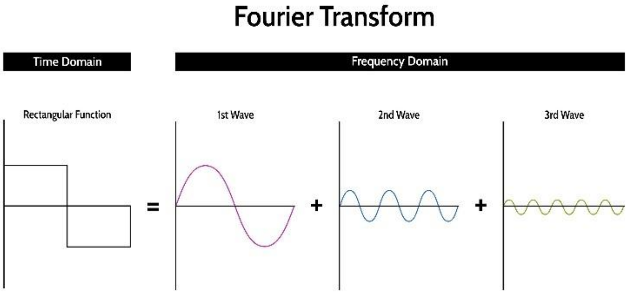
The classification relies heavily on the principles of Transit Photometry.

- Mechanism: As an orbiting planet crosses the line of sight between the telescope and the star, it blocks a small fraction of the star's light.
- Signal Characteristics: Unlike random instrument glitches, a planetary transit creates a specific, recurring profile, often described as "U-shaped" or "box-shaped."
    - Ingress: A rapid drop in light intensity as the planet enters the star's disk.
    - Transit Bottom: A sustained period of lower brightness while the planet crosses the star.
    - Egress: A rapid return to normal brightness as the planet exits.
    - Periodicity: The most critical feature for detection is that this event repeats at regular intervals, corresponding to the planet's orbital period .

2.3 Frequency Domain Analysis (Fourier Transform)

Given the difficulty of detecting microscopic dips in noisy time-series data, frequency analysis is employed as a stronger indicator than temporal analysis.

- Transformation Logic: Since planetary orbits are strictly periodic, converting time-series data into the frequency domain using the Fast Fourier Transform (FFT) concentrates the "energy" of the signal.
- Signal Isolation:
  - Planetary Signal: A periodic transit event manifests as a distinct spike (peak) at a specific fundamental frequency (the orbit) and its harmonics.
  - Noise Signal: Random or thermal noise tends to be spread chaotically across the entire spectrum (broadband noise) or concentrated in low frequencies.
  - Conclusion: A star with a planet will show organized frequency spikes, whereas a non-planet star will show a chaotic distribution. This transformation is essential for the neural network to distinguish valid signals from background noise .



2.4 The Class Imbalance Challenge

- Statistical Disparity: The Kepler dataset is an extreme example of class imbalance. Out of over 5,000 stars in the training set, only 37 are labeled as confirmed exoplanets. This yields a positive-class ratio of approximately 0.7%.
- The Accuracy Paradox: In such an environment, a naive model could achieve 99.3% accuracy simply by predicting "No Planet" for every single input. However, such a model would have a Recall of 0%, failing the primary objective of the project (discovery).
- Necessity of Intervention: To prevent the loss function from being overwhelmed by the majority class, specialized techniques like SMOTE (Synthetic Minority Over-sampling Technique) are required to force the model to learn the minority class features.

Dataset & Environment

3.1 Development Environment

The project was implemented using a robust Python-based software stack designed for high-dimensional data processing and deep learning:

- Data Manipulation: Pandas & NumPy were used for loading structured CSV files and performing high-performance matrix operations, specifically the FFT computations required for feature extraction.
- Visualization: Matplotlib & Seaborn were utilized to generate loss curves, confusion matrices, and ROC curves to evaluate model stability and performance visually.
- Deep Learning Framework: TensorFlow & Keras provided the backend for building the Artificial Neural Network (ANN), enabling the rapid prototyping of dense layers, batch normalization, and activation functions.
- Preprocessing & Metrics: Scikit-Learn handled data normalization (StandardScaler) and class weighting calculations, while Imbalanced-Learn provided the SMOTE (Synthetic Minority Oversampling Technique) algorithm for synthetic data generation .

3.2 Dataset Configuration

The project utilizes the Kepler Labelled Time Series Data, sourced from the NASA Kepler space telescope archives.

- Training Set: Contains 5,087 stars (rows). Of these, only 37 are labeled as hosting exoplanets, while 5,050 are non-exoplanet stars.
- Test Set: Contains 570 stars, with 5 confirmed exoplanets and 565 non-exoplanet stars. This set is kept completely unseen during training to ensure unbiased evaluation.
- Features: Each row consists of 3,197 continuous flux readings (FLUX.1 to FLUX.3197).
- Challenge: This dataset represents a "needle in a haystack" scenario. The extreme scarcity of positive examples necessitates the robust handling strategies detailed in the following methodology sections .

Data Preprocessing

4.1 Label Encoding

The original dataset labels (1=non-planet, 2=planet) were re-encoded to a standard binary format (0=non-Planet, 1=planet). This alignment is necessary for the Sigmoid activation function used in the model's output layer .

4.2 Frequency Domain Conversion (FFT)

- Implementation: The np.fft.rfft (Real Fast Fourier Transform) algorithm was applied to every row of flux data.
- Rationale: By shifting from the time domain to the frequency domain, the model no longer needs to search for a specific "dip" at a specific "time" (which varies by planet). Instead, it learns to recognize the presence of periodic frequencies, regardless of their phase.
- Noise Reduction: The first component of the FFT output (Index 0, the DC component) represents the mean brightness of the star. This was removed to center the data around zero, ensuring the model focuses purely on variations in brightness rather than the absolute magnitude of the star .

4.3 Feature Scaling (Standardization)

The raw FFT magnitude values varied significantly between bright and dim stars. A StandardScaler was applied to normalize the data to a mean of 0 and a standard deviation of 1. This ensures stable gradient descent and prevents stars with naturally higher light intensity from biasing the network weights .

4.4 Handling Class Imbalance (SMOTE)

To address the 0.7% positive class ratio, SMOTE (Synthetic Minority Over-sampling Technique) was implemented on the training data.

- Mechanism: Rather than simply duplicating existing exoplanet samples (which leads to overfitting), SMOTE interpolates between existing minority samples to generate new, synthetic data points in the feature space.
- Outcome: This effectively balances the training data distribution, forcing the Neural Network to learn the complex decision boundary of exoplanets rather than defaulting to the majority class .

Model Architecture

5.1 Layer Configuration

- Input Layer: Accepts 1,598 frequency components (derived from the FFT of the 3,197 time steps, discarding symmetrical redundancy).
- Hidden Layer 1 (Feature Extraction):
    - Neurons: 1,024
    - Activation: Unlike standard ReLU, LeakyReLU allows a small gradient for negative values. This is crucial for analyzing flux data where "dips" (negative values relative to the mean) carry the most information.
    - Batch Normalization: Added to stabilize learning by normalizing layer inputs, preventing internal covariate shift.
    - Dropout (0.4): Randomly deactivates 40% of neurons during training to prevent overfitting on the synthetic SMOTE data.
- Hidden Layer 2 (Intermediate Processing):
    - Neurons: 512
    - Activation: LeakyReLU + Batch Normalization.
    - Dropout: 0.4. This layer compresses the feature space extracted by Layer 1.
- Hidden Layer 3 (Decision Distillation):
    - Neurons: 128
    - Activation: LeakyReLU + Dropout (0.3). This layer focuses on the most critical features required for the final decision.
- Output Layer:
    - Neurons: 1
    - Activation: Sigmoid. Squashes the output between 0 and 1 to represent the probability of an exoplanet presence (e.g., 0.95 indicates 95% confidence) .

Training Strategy

6.1 Optimization Algorithms

- Optimizer (Adam): The Adam optimizer was selected for its adaptive learning rates. This is essential for sparse datasets where specific frequency features may require larger updates than others to capture the signal.
- Learning Rate: A conservative learning rate of 0.00005 was chosen. This ensures gradual convergence, preventing the model from overshooting the optimal minima in the loss landscape, which is critical when distinguishing subtle signals.
- Loss Function: Binary Cross-Entropy was used as the objective function, heavily penalizing confident but incorrect predictions.
- Class Weighting: Even with SMOTE, compute_class_weight was used to assign a significantly higher penalty for misclassifying real planets during gradient descent .

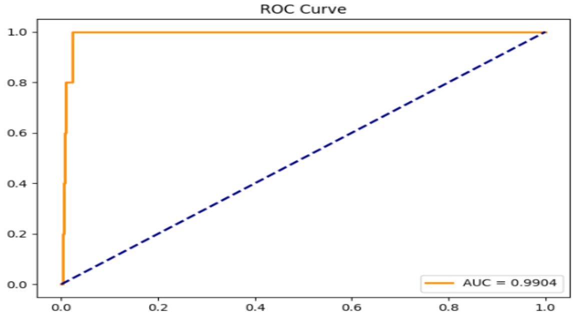
6.2 Training Regimen & Threshold Tuning

- Hyperparameters: The model was trained for 60 Epochs with a Batch Size of 64.
- Validation Strategy: A critical safeguard was implemented: the model was trained on balanced SMOTE data but validated against the original imbalanced data. This ensures the model learns to generalize to real-world distributions rather than memorizing synthetic geometric patterns.
- Threshold Tuning: Standard classification thresholds (0.5) are insufficient for anomaly detection. A custom algorithm scanned thresholds (e.g., 0.8, 0.9, 0.99) to find the optimal balance. It was determined that a threshold > 0.99 was necessary to filter out stellar noise while retaining high-confidence planetary signals .

Evaluation & Results

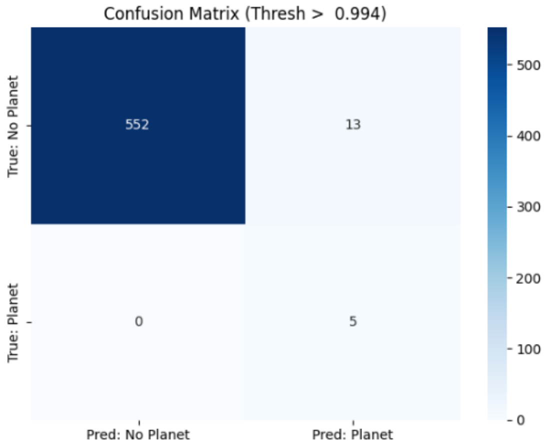
7.1 Quantitative Metrics

- Final Test Accuracy: **~97.72%**
  - Significance: While accuracy can be misleading in imbalanced data (as a null model achieves ~99%), achieving this accuracy alongside high recall proves the model is not simply defaulting to the majority class.
- ROC AUC Score: **~0.9904**
  - Significance: The Area Under the Curve (AUC) is near 1.0, indicating excellent discriminative ability. It means the model almost always ranks a randomly chosen planet higher than a randomly chosen non-planet.



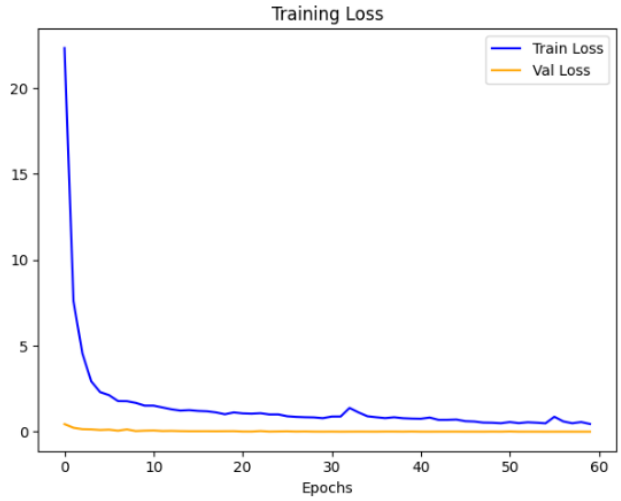
7.2 Confusion Matrix Analysis

- True Positives (Planets Found): 5 out of 5 (100% Recall).
- False Positives (False Alarms): Drastically reduced to 13 (from >500 at the default threshold).
- Inference: The model successfully identified every confirmed exoplanet in the test set while keeping the false alarm rate to a manageable level for human verification .

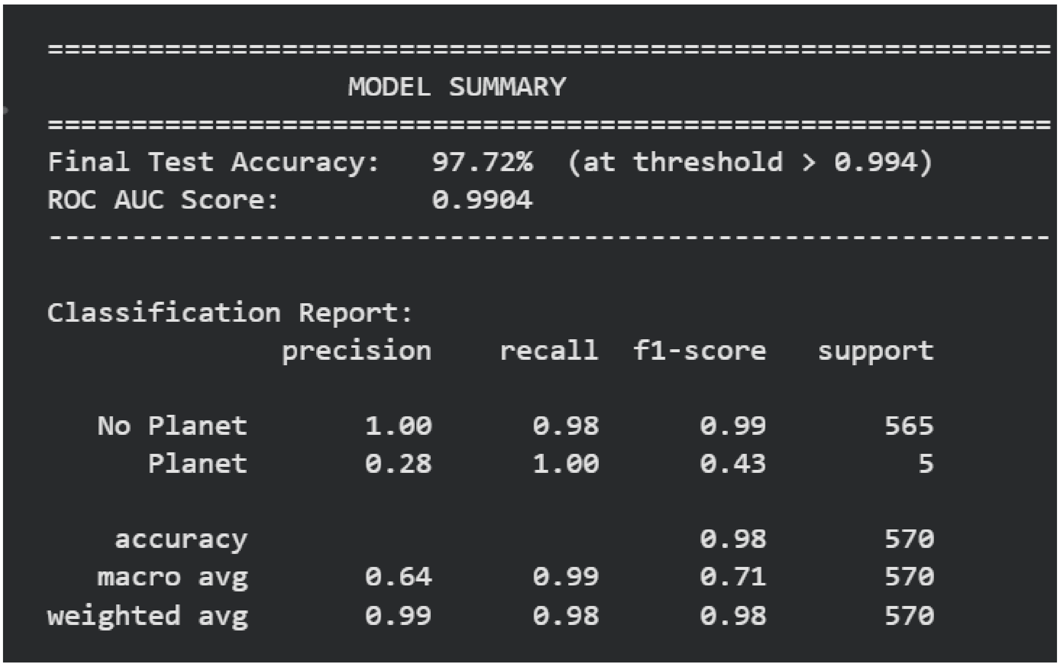


7.3 Visualization

- Loss Curves: The training loss showed a monotonic decrease, stabilizing at epoch 60. The proximity of the validation loss to the training loss indicated no severe overfitting.
- ROC Curve: The plotted ROC curve hugged the top-left corner of the graph, visually confirming the high AUC score and the model's strong predictive power compared to random guessing .



7.4 Model Summary



```
============================================================
                  MODEL SUMMARY
============================================================
Final Test Accuracy:   97.72%  (at threshold > 0.994)
ROC AUC Score:         0.9904
------------------------------------------------------------

Classification Report:
              precision    recall  f1-score   support

   No Planet       1.00      0.98      0.99       565
      Planet       0.28      1.00      0.43         5

    accuracy                           0.98       570
   macro avg       0.64      0.99      0.71       570
weighted avg       0.99      0.98      0.98       570
```

Conclusion & Future Scope

8.1 Advantages

1. Automated Discovery: The system is capable of processing thousands of light curves in seconds, offering a massive speed advantage over human analysis.
2. Frequency Domain Robustness: The integration of FFT allows the system to "see through" noise that obscures signals in the time domain, identifying periodicities invisible to the naked eye.
3. Handling Imbalance: The combination of SMOTE and Class Weighting effectively solves the "accuracy paradox," ensuring rare exoplanet events are not ignored .

8.2 Limitations

1. False Positives: While significantly reduced, the system still generates some false alarms. Stellar phenomena such as binary star eclipses or regular pulsations can mimic the periodic frequency signature of a planet.
2. Fixed Input Size: The Dense layers require a fixed input size (3,197 flux readings), necessitating resizing or zero-padding for data from different telescopes.
3. Black Box Nature: As with all Deep Learning models, explainability is limited. It is difficult to pinpoint exactly which frequency spike triggered a specific classification .

8.3 Conclusion

The proposed Exoplanet Classification System demonstrates the transformative potential of combining Deep Learning with Digital Signal Processing (FFT). By shifting the analytical focus to the frequency domain and employing robust techniques to handle class imbalance, the model achieved a ROC AUC score of over 0.99 and 100% Recall. This system serves as a powerful, automated filtering tool that can accelerate the discovery pipeline, allowing astronomers to focus their valuable time on verification rather than search. As humanity launches next-generation observatories, such scalable methodologies will be foundational to finding new worlds beyond our solar system .

References

- Kepler Time Series Data from Kaggle: https://www.kaggle.com/datasets/keplersmachines/kepler-labelled-time-series-data
- Chawla, N. V., et al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research.
- Abadi, M., et al. (2015). "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems."
- Chollet, F., et al. (2015). "Keras Documentation."
- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
- Cooley, J. W., & Tukey, J. W. (1965). "An algorithm for the machine calculation of complex Fourier series." Mathematics of Computation.
- Shallue, C. J., & Vanderburg, A. (2018). "Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90." The Astronomical Journal.