The Identification of Exoplanets using Physics-Aware Machine Learning

Phase II: Model Implementation and Results

Esraaj Sarkar Gupta Suyash Prakash Anirudh Puri

May 2026

Machine Learning and Pattern Recognition | Plaksha University

The Project: Beyond Black-Box Models

Core Objective:
We aim to identify and validate exoplanets by integrating fundamental orbital mechanics into classical machine learning architectures.

Our Thesis:
Physics-Informed Machine Learning (PIML) provides superior interpretability, robustness, and performance over purely data-driven models. By engineering physical boundaries, we dramatically reduce sample complexity.

> *"Early results show that relatively simple ML techniques paired with well-engineered features often perform comparable to much more complex ML models... critical for establishing credibility."* [1]

[1] Ford, E. B. (2025). Enhancing Exoplanet Surveys via Physics-informed Machine Learning. *Proc. of the IAU.*

Mission Overview & Datasets

Our analysis leverages data from three distinct phases of space-based transit photometry.[2]

Kepler DR25 (KOI)

The Gold Standard

Four years of continuous monitoring in a fixed field. Provided the highly stable, complete cases used to train our physical anchors.

K2 Mission

Kepler Extended

Born from the mechanical failure of Kepler's reaction wheels. Surveyed the ecliptic plane using solar radiation pressure for stability.

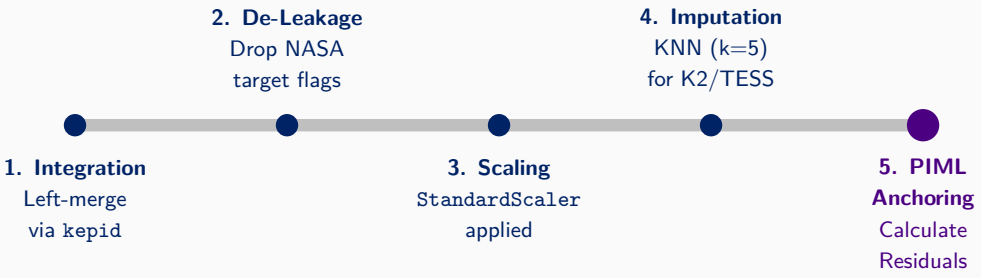
TESS

Current Survey

An all-sky scanning mission focusing on stars 30-100x brighter than those in the Kepler field. Exhibits high sparsity.

[2] Data sourced from the NASA Exoplanet Archive: `https://exoplanetarchive.ipac.caltech.edu/`

Data Preprocessing Pipeline



Handling Sparsity: Kepler data was robust enough for complete-case analysis. However, TESS and K2 exhibited severe observational gaps (nearly every row contained a `NaN`). Distance-weighted KNN imputation was vital to preserve physical covariance structures before feature engineering.

Physics-Informed Feature Engineering

To constrain our machine learning models, we explicitly calculate the residuals (Δ) between the telescope's observations and theoretical orbital mechanics.

- **1. Transit Depth Consistency** ($\Delta\delta$)

$$\delta_{\text{theo}} = \left(\frac{R_p}{R_*}\right)^2 \quad \Longrightarrow \quad \Delta\delta = |\delta_{\text{obs}} - \delta_{\text{theo}}|$$

- **2. Transit Duration Consistency** (ΔT)

$$T_{\text{theo}} = \frac{P}{\pi} \arcsin\left(\frac{R_* \sqrt{1-b^2}}{a}\right) \quad \Longrightarrow \quad \Delta T = |T_{\text{obs}} - T_{\text{theo}}|$$

- **3. Impact Parameter Consistency** (Δb^2)

$$b^2_{\text{theo}} = \left(1 - \sqrt{\delta}\right)^2 - \left(\frac{a}{R_*} \sin\left(\frac{T\pi}{P}\right)\right)^2 \quad \Longrightarrow \quad \Delta b^2 = |b^2_{\text{obs}} - b^2_{\text{theo}}|$$

Validating Theoretical Anchors

Separation of Classes via Physical Residuals: Before passing data to the model, we verified that our engineered physics residuals naturally isolate False Positives (red) from Confirmed Exoplanets (green).

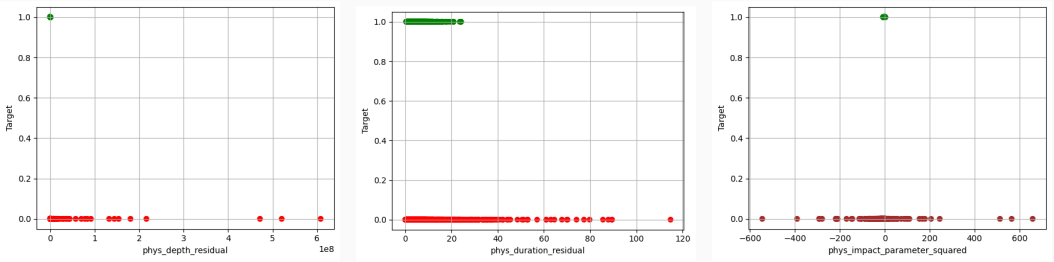
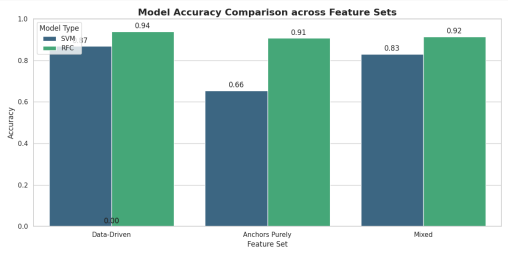


Figure 1: Distributions of residuals for Transit Depth, Duration, and Impact Parameter (b^2). Confirmed planets tightly hug zero (theoretical agreement), while false positives cascade outward.

Model Architecture Selection: SVM vs. RFC

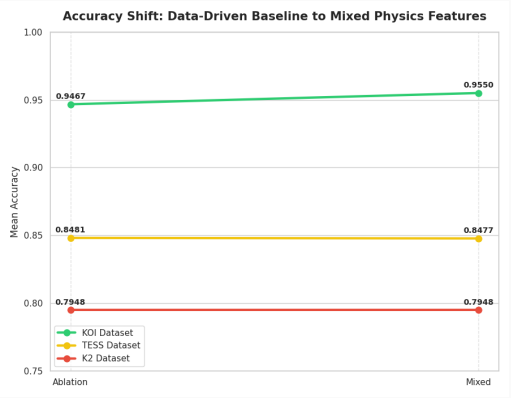
To capture complex, non-linear relationships without the overhead of deep learning, we initially selected Support Vector Classifiers (SVC) and Random Forest Classifiers (RFC) as our core baselines.

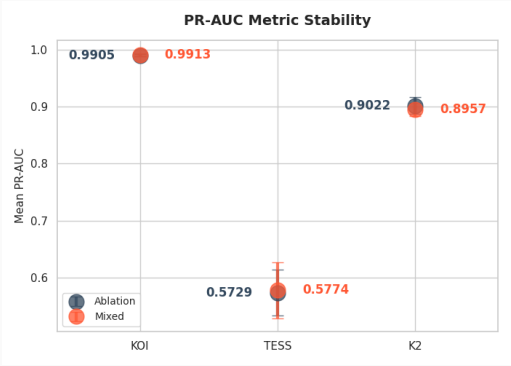


RFC significantly outperformed SVM across all folds.

Ablation Study: Data-Driven vs. Physics-Informed (RFC)

We conducted a head-to-head ablation study comparing a purely Data-Driven approach vs. our Best PIML framework.





Ablation Insight: The Impact of Non-Gaussian Noise

The K2 Anomaly Explained:

Why did our purely data-driven model outperform the physics-anchored framework specifically on the K2 dataset?

K2 Dataset Performance	Accuracy	PR-AUC
Data-Driven (Ablation Set)	**0.8380**	**0.9375**
Physics-Informed	0.7879	0.8855

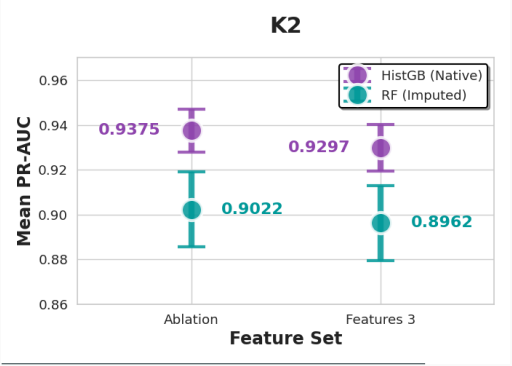
Recent literature confirms that Physics-Informed models degrade significantly under non-Gaussian noise.[3] Rigid physical equations force the model to fit to flawed theoretical assumptions, whereas unconstrained data-driven models adaptively learn and bypass the noise distribution.

[3] Pilar, P., & Wahlström, N. (2023). Physics-informed neural networks with unknown measurement noise. *NeurIPS 2023 Workshop on Machine Learning and the Physical Sciences*.

Handling Sparsity: HistGB vs. KNN Imputation

What is HistGradientBoosting (HistGB)?

It is a highly efficient, histogram-based tree architecture. Crucially, unlike Random Forests, HistGB [4] **natively handles missing values (NaNs)** by learning during training which child node the missing data should be routed to.



[4] Ke, G., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS 2017*. (The basis for Scikit-Learn's HistGradientBoosting).

Comparisons with Deep Learning: TabNet

What is TabNet?

Developed by researchers at Google Cloud AI, TabNet[5] is a deep learning architecture designed specifically for tabular data. It uses a sequential attention mechanism to choose which features to reason from at each decision step.

[5] Arik, S. O., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. *AAAI Conference on Artificial Intelligence*.

Model Explainability: Opening the "Black Box"

A core advantage of Classical ML over Deep Learning is the retention of **Explainability**. Using Mean Gini Importance, we can quantify exactly which features drive the model's decisions.

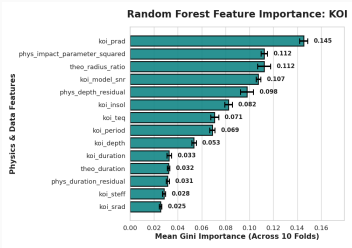


Figure 2: Kepler (KOI)

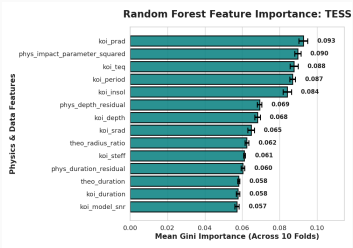


Figure 3: TESS

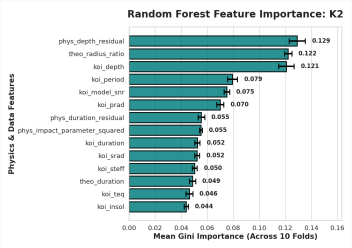


Figure 4: K2

Questions?

Thank You!

Esraaj Sarkar Gupta • Suyash Prakash • Anirudh Puri
Machine Learning and Pattern Recognition | Plaksha University