IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Automated Exoplanet Detection Using Artificial Intelligence and Machine Learning

Jiya Kishor Patel[1], Isha Sachin Jadhav[2], Sakshi Madhukar Bari[3], Sakshi Mangeshrao Patil[4]

[1]*Engineering Student, Department of Computer Engineering, R. C. Patel Institute of Technology Shirpur-425405, Dist-Dhule, Maharashtra, India*

[2]*Engineering Student, Department of Computer Engineering, R. C. Patel Institute of Technology Shirpur-425405, Dist-Dhule, Maharashtra, India*

[3]*Engineering Student, Department of Computer Engineering, R. C. Patel Institute of Technology Shirpur-425405, Dist-Dhule, Maharashtra, India*

[4]*Engineering Student, Department of Computer Engineering, R. C. Patel Institute of Technology Shirpur-425405, Dist-Dhule, Maharashtra, India*

Abstract – *Exoplanet research has grown into one of the most compelling frontiers in modern astrophysics reshaping what we know about worlds beyond our own solar system. These are planets orbiting stars other than the Sun and detecting them is rarely straightforward, astronomers typically rely on indirect methods, with transit photometry being among the most productive [1]. Space missions like the Kepler Space Telescope and TESS have generated enormous quantities of observational data, collectively responsible for confirming thousands of exoplanet candidates [3][4].Even so, a large share of that data has historically required manual review, which is both slow and inconsistent in practice. Classical detection tools like the Box Least Squares algorithm do a reasonable job of flagging periodic dips in light curves, but they struggle with noise and frequently flag non-exoplanets as real candidates. This paper introduces a hybrid detection framework that draws on traditional algorithms alongside machine learning and deep learning methods. Combining Random Forest classifiers, convolutional neural networks and oversampling via SMOTE, the system targets three key improvements: fewer false positives, stronger detection accuracy and the ability to scale across much larger observational archives.*

Keywords - exoplanet detection, machine learning, deep learning, light curve analysis, space exploration, transit photometry.

INTRODUCTION

The discovery of a Jupiter-sized companion orbiting a neighboring sun-like star in 1995, which demonstrated that planetary systems are not exclusive to our region of the galaxy, is probably the beginning of modern exoplanet study [2]. The breadth and ambition quickly expanded as a result. Increased observational activities and space missions were created especially for photometry in transit. For years, the Kepler telescope observed the same region of the sky, documenting changes in the brightness of tens of thousands of stars and keeping an eye out for the recurring dips that indicate a planet in orbit [3]. By scanning various parts of the sky and concentrating on stars close enough for in-depth spectroscopic follow-up, K2 and TESS expanded that work [4].Although the scientific results have been astounding, the data problem that accompanied them is just as important. Thousands of light curves are produced by each mission, and it is not feasible nor reliable to manually sort through that volume. Particularly when signals are weak, erratic, or buried in noise from star activity or instrumental drift, human assessors tend to miss things. The gap between what is collected and what can be realistically investigated only gets wider as mission archives continue to expand, making it difficult for traditional analytic processes to keep up. In this context, automated detection is essential rather than optional. By immediately learning pattern recognition from labeled samples and substituting hand-crafted rules with representations that adapt to the data.Purely rule-based approaches are ill-suited to deal with the richness and diversity of contemporary astronomical datasets.

The morphology of a true transit can differ significantly based on planet size, orbital inclination, and host star characteristics; different missions employ different cadences; and different stars generate distinct noise profiles. Compared to fixed algorithms, learned models are more adaptable in navigating this heterogeneity, identifying feature combinations that are challenging to predetermine. The speed gains are also important; automating the identification step makes processing big information in real time much more practical.

These techniques' capacity to distinguish genuine transits from stellar noise and instrumental artifacts is further refined by grounding them in astrophysical context and utilizing physically significant features in addition to raw waveforms. When a system uses both of these inputs, it can see the situation more clearly than when it uses only one. In the future, the argument for this type of data-driven, adaptive detection will only get stronger. The archive is being expanded by missions, and any framework that is unable to keep up with fresh observations would soon lag behind. The field needs a system that scales without requiring continuous human intervention, functions consistently under a variety of observational situations, and improves gradually as data accumulates. This effort is intended to solve the issue.

LITERATURE REVIEW

Over the past 20 years, there has been a significant movement in the methods used to detect exoplanets, from simply statistical techniques to more advanced computational tools. The most effective detection method still in use is transit photometry, which searches for recurring, distinctive decreases in a star's brightness as a planet passes in front of it. In order to locate the best-fitting periodic dip, Box Least Squares was designed to search light curves for precisely these signatures, examining numerous potential periods and lengths [6]. It performs well in clean settings, but it produces a significant amount of false detections due to noisy data and stellar fluctuation. By including methodical screening procedures, the Kepler data processing pipeline expanded on this, generating sorted lists of transit candidates that could then be classified using machine learning techniques. Because Random Forest can handle a large number of input features without overfitting and has continuously demonstrated strong performance across several exoplanet benchmark datasets, it proved to be a good choice for this classification problem [9].

Random Forest has continuously demonstrated good performance across several exoplanet benchmark datasets and is well-suited to this classification challenge due to its ability to handle huge quantities of input features without overfitting [9]. These improvements were validated by later research using various dataset configurations, with supervised classifiers achieving accuracy levels far higher than those of conventional techniques [7]. By working directly on raw time-series waveforms instead of pre-extracted information, convolutional networks advanced this.One well-known example is Astronet, which showed that a trained CNN could consistently distinguish between a variety of astronomical false positives and real planetary transits [5]. Class imbalance is a recurrent problem in all of these methods: classifiers are biased toward the majority class because confirmed exoplanets make up a very small fraction of star observations.In order to overcome this, SMOTE creates artificial minority-class samples during training. Research employing this method has shown significant increases in recall without compromising precision [10]. Similar advantages have been demonstrated by work on synthetic data augmentation more generally [8].

OBJECTIVES

- To create an automated exoplanet finding system that can effectively handle massive astronomical datasets by utilising machine learning and artificial intelligence approaches.
- To increase classification performance by combining machine learning and deep learning models with conventional techniques to increase detection accuracy.
- To lower false positive rates by improving signal identification through efficient preprocessing and feature engineering methods.
- To use appropriate data balancing and optimisation techniques to address class imbalance and guarantee scalability for huge astronomical datasets.

DATASET DESCRIPTION

Light curve records from the Kepler and TESS missions, which are publicly accessible NASA archives ,provided the photometric data used in this study [11] .These databases comprise labelled observations that identify verified exoplanets and false positives, and they monitor changes in star brightness over time. Each light curve has thousands of flux measurements, as shown in Fig. 1 ,where a planetary transit manifests as a recurring drop in an otherwise steady signal.The significant class imbalance in this data, where true exoplanet detections are far less common than non-detections, is a major problem. The preprocessing pipeline uses SMOTE-based oversampling, feature scaling and flux normalisation to address this [10]. Signal-to-noise ratio, orbital period and transit depth and duration are important characteristics.

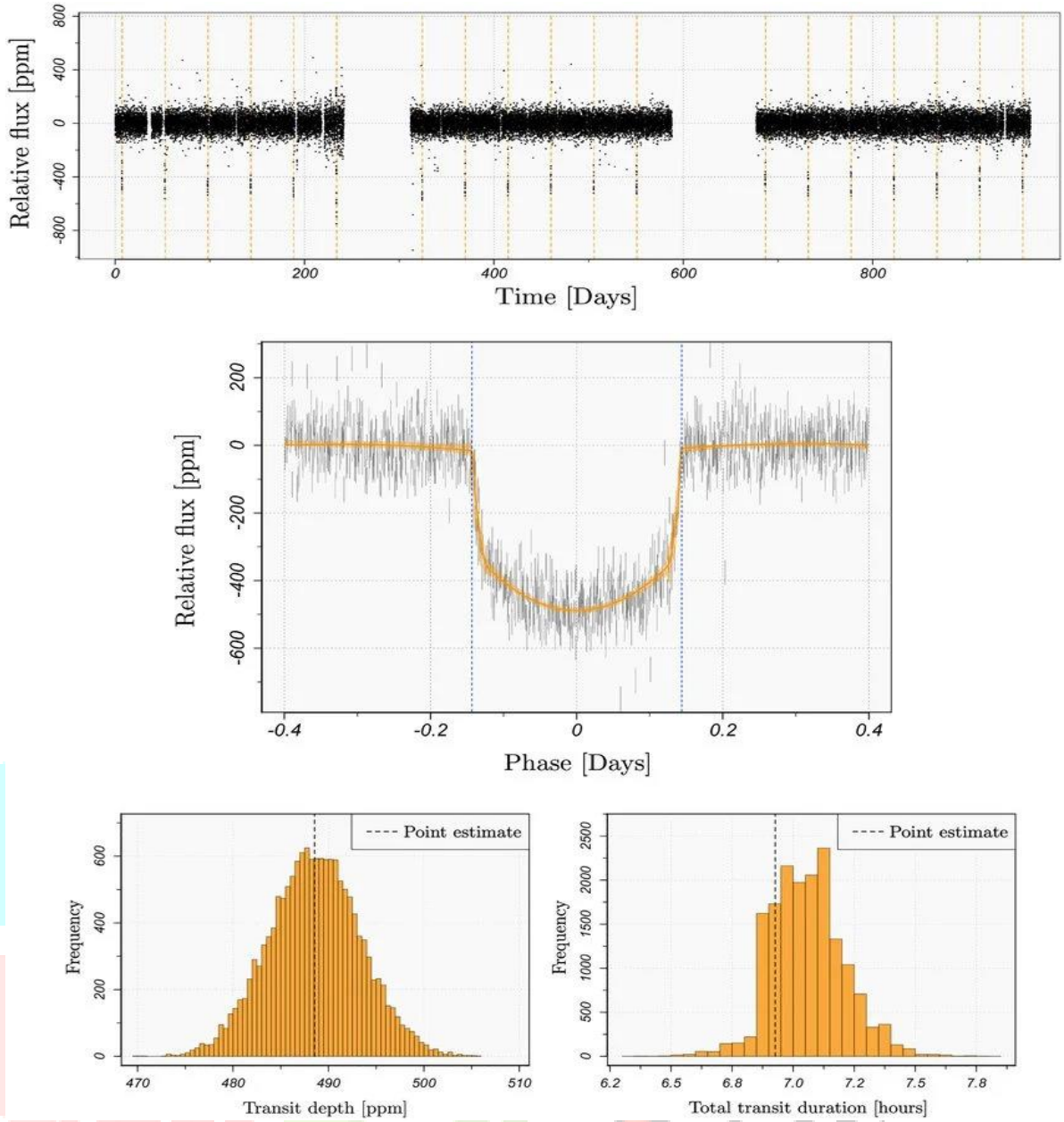

Fig. 1 - Representative photometric light curve from the NASA Kepler archive. The periodic "dips" in stellar flux indicate a candidate planetary transit against background noise.

METHODOLOGY

Using Kepler and TESS time-series photometry, the detection pipeline starts with gathering raw data from NASA's publicly accessible sources. Every record is a light curve which is a timestamped series of flux measurements that show how a star's brightness changes across the observation frame. First, information is taken from the Transiting Exoplanet Survey Satellite and Kepler missions. A timestamped series of relative flux measurements collected over months or years of continuous observation makes up each light curve. The data undergoes a cleaning stage prior to any modeling step: outliers from stellar flare events or instrumental glitches are eliminated, flux readings are standardized to enable comparisons between stars and observation periods, and missing values are filled using imputation. At this point, the shorter-timescale dips that are the real detection target are separated from long-duration brightness patterns that are unrelated to transits. The next step is feature extraction, which generates the numerical inputs that classifiers will utilize.

Together, transit depth, duration, orbital period, and signal-to-noise ratio provide the model with sufficient data for classification. Each of these parameters carries diagnostic information about whether a particular dip is most likely planetary in origin. In order to assess how well the trained models generalize to new observations, the labeled dataset is subsequently divided into training and held-out test sets. There are three steps in the detection process. BLS is first used as a general search, looking for recurring periodic signals in each light curve and marking the best candidates for additional examination. In order to effectively and comprehensibly filter out a significant percentage of false positives, Random Forest then uses the collected characteristics for those flagged candidates to assign classification probabilities. In order to identify morphological features that the feature-based classifier might have overlooked, a convolutional neural network then analyzes the raw waveform segments around the highest-probability

candidates. With each stage making up for the failure modes of the others, the combination of all three stages performs better than any one approach alone. .

ALGORITHMS USED

In order to determine which configuration best fits the observed flow pattern, BLS tests a grid of candidate periods and transit durations, fits a rectangle model to the data at each combination, and then chooses that configuration. When stellar variability provides confounding patterns or when noise levels are high, it significantly deteriorates. However, it is dependable for clean, periodic signals. Four classifiers are used in the machine learning layer: Support Vector Machine, Random Forest, Decision Tree, and Logistic Regression. To distinguish true planetary signals from false positives, each analyses the identical collection of extracted transit properties. Random Forest typically outperforms the others because of its ensemble structure, which is composed of numerous independent decision trees that vote on each classification. This structure naturally resists overfitting and provides good generalisation over a variety of feature combinations [9]. Convolutional neural networks take over and analyse raw light curve segments directly at the deep learning stage. CNNs create their own hierarchical representations from the waveform data instead of depending on pre-defined features, learning how a transit appears at various sizes and noise levels. Because of this, they are particularly useful for recovering detections that don't fit cleanly into the feature space used to train the previous classifiers.

RESULTS AND EVALUATION

Five common metrics were used to evaluate the system's performance: accuracy, precision, recall, F1-score, and area under the ROC curve. Recall is especially important in the situation of exoplanet detection; a missed real discovery has more repercussions than a false alarm that is eliminated during follow-up. To make sure the generated candidate lists continued to be scientifically valuable, precision and F1-score were monitored in addition to recall. The machine learning models significantly outperformed traditional-only baselines across the test set. When compared to BLS-only detection, misclassification rates on structured data significantly decreased, and classifiers were able to distinguish between planetary and non-planetary signals more precisely thanks to the richer feature representations. The deep learning model produced the best overall results, as seen in Fig. 2, especially for low-amplitude events and ambiguous transit shapes where the signal does not stand clearly above the noise floor. CNNs' advantage in these more difficult scenarios can be explained by their ability to capture non-linear patterns in raw flux sequences that are difficult to quantify cleanly in any predefined feature space [5].Across all parameters, the whole hybrid pipeline yielded the greatest outcomes, demonstrating that the three steps are actually complementary rather than redundant. BLS maintains a wide initial search; Random Forest effectively reduces false positives by employing structured features; and CNN retrieves edge instances that were previously discarded.Additionally the pipeline performed well when evaluated on bigger data partitions, indicating that it can generalize to complete mission archives without requiring a significant amount of retraining.

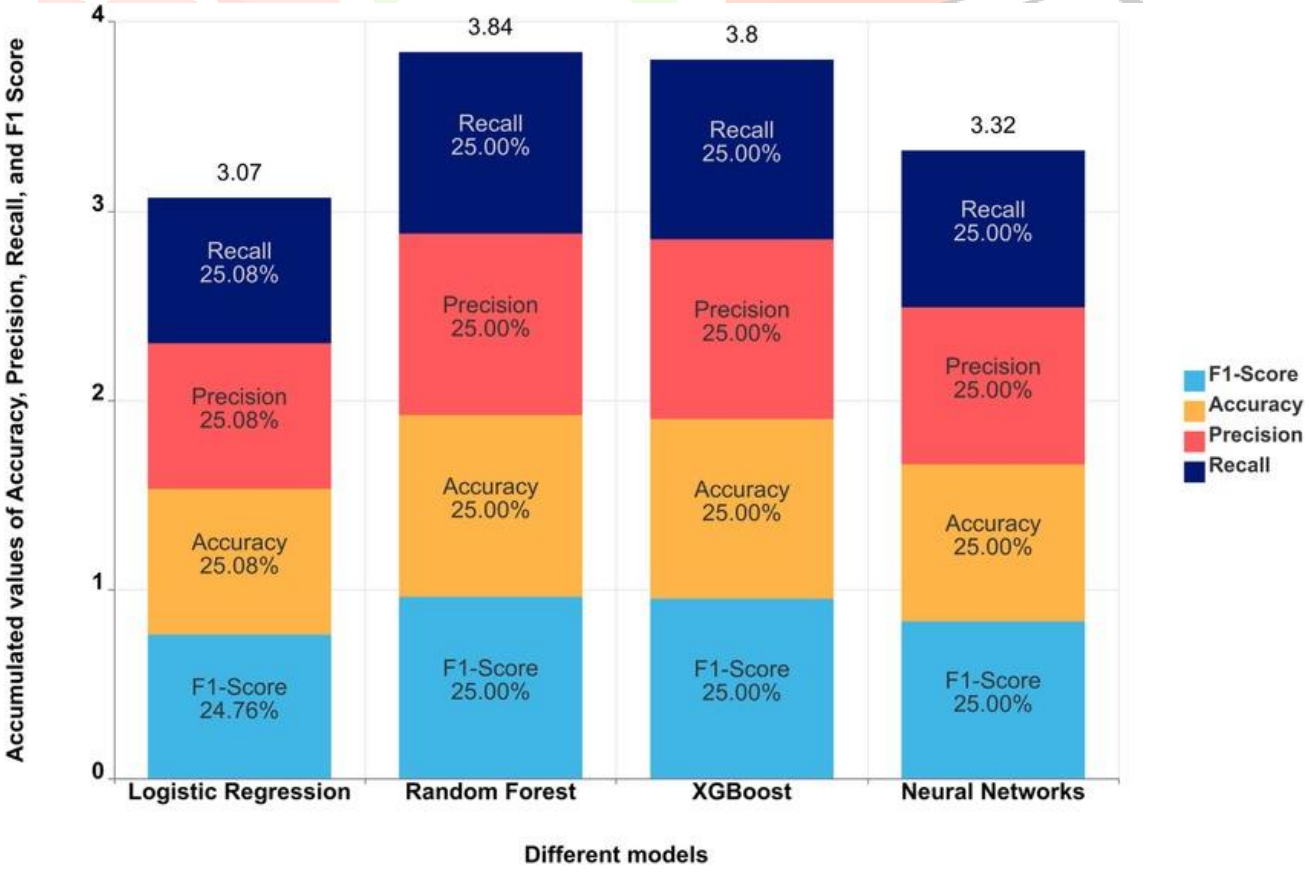

Fig. 2 - Comparative analysis of model performance. The hybrid framework demonstrates superior Accuracy and Recall metrics relative to individual machine learning baselines.

CONCLUSION

This paper developed and assessed a three-stage detection pipeline that combines convolutional networks , Random Forest classification and BLS into a framework where each element tackles the failure modes of the others. The findings verify that the hybrid technique addresses class imbalance using SMOTE-based oversampling, grows to bigger data divisions without degradation, and exceeds single-method baselines in accuracy and recall. In terms of practical impact ,reducing the amount of human labour required to find qualified follow-up candidates is the most immediate benefit; this issue becomes more urgent as mission archives develop. Adapting the pipeline for streaming data ingestion during active mission operations and adding multi-band or spectroscopic inputs to improve the distinction between actual transits and stellar activity are the remaining open directions. The architecture employed here is built to grow with the far larger and more varied archives that future missions will bring. The overarching goal is to develop a detection system that keeps up with the data, highlighting the most intriguing candidates from an ever-expanding observational recole astronomical dataset. On the held-out test set, the pipeline continuously surpassed single-method baselines in accuracy and recall. The improvements persisted when assessed on bigger data partitions, which is a fair proxy for deployment on whole mission archives. The stages of machine learning and deep learning turned out to be complementary: convolutional networks recover edge instances that defy neat feature-based description, while supervised classifiers effectively minimise false positives on structured data.In the future, a number of extensions are worthwhile. Instead of being limited to batch post-processing, real-time ingestion of incoming data would enable the system to be used during actual mission operations. The ability to distinguish between real transits and stellar imposters could be improved by including stellar parameters or multi-band photometric observations as inputs. Because of the pipeline's modular design, it is possible to upgrade specific components as new algorithms become available without having to rebuild the entire system. The ultimate objective is a detection infrastructure that continuously surfaces the highest-priority candidates for scientific follow-up, keeps up with the observation rates of ongoing and prospective missions, and becomes more dependable as its training data develops.

REFERENCES

1. *NASA Exoplanet Exploration Program, "What is an Exoplanet?" [Online]. Available: https://exoplanets.nasa.gov/faq/3/what-is-an-exoplanet/*
2. *M. Mayor and D. Queloz, "A Jupiter-mass companion to a solar-type star," Nature, 1995.*
3. *W. J. Borucki, "Kepler Mission Overview," [Online]. Available: https://arxiv.org/abs/1001.0268*
4. *G. R. Ricker et al., "Transiting Exoplanet Survey Satellite," J. Astron. Telesc. Instrum. Syst., 2015. [Online]. Available: https://doi.org/10.1117/1.JATIS.1.1.014003.*
5. *C. J. Shallue and A. Vanderburg, "Astronet: Deep Learning for Exoplanets," [Online]. Available: https://arxiv.org/abs/1712.05044*
6. *J. M. Jenkins et al., "Kepler Data Processing Pipeline," [Online]. Available: https://arxiv.org/abs/1001.0258*
7. *Karimi et al., "Machine Learning for Exoplanet Detection," [Online]. Available: https://arxiv.org/abs/2508.09689*
8. *E. Lo, "Synthetic Data for Exoplanet Detection," [Online]. Available: https://arxiv.org/abs/2507.19520*
9. *L. Breiman, "Random Forest," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.*
10. *N. V. Chawla et al., "SMOTE Technique," [Online]. Available: https://www.jair.org/index.php/jair/article/view/10302*
11. *NASA Exoplanet Archive, "Kepler and K2 Time Series Data," NASA, 2024. [Online]. Available: https://exoplanetarchive.ipac.caltech.edu/*
12. *C. J. Shallue and A. Vanderburg, "Identifying Exoplanets with Deep Learning," The Astronomical Journal, vol. 155, no. 2, 2018*