ACRONYMS

ACRONYM	FULL FORM
AEC	Ability Enhancement Course
CNN	Convolutional Neural Network
ML	Machine Learning
NASA	National Aeronautics and Space Administration
SVM	Support Vector Machine
TPS	Transit Planet Search
VTU	Visvesvaraya Technological University

CHAPTER 01

INTRODUCTION

CHAPTER 01

INTRODUCTION

1.1 About the Problem Area

The discovery of exoplanets i.e., planets orbiting stars beyond our Solar System has transformed modern astronomy and astrophysics. Over the last few decades, advances in space-based observations and data analysis techniques have enabled scientists to detect thousands of such planets, revealing that planetary systems are common in our galaxy. Among the various detection techniques, **transit photometry** has emerged as one of the most successful and widely used methods.

NASA's Kepler Space Telescope played a pivotal role in this progress by continuously monitoring the brightness of more than 150,000 stars. When a planet passes in front of its host star, it causes a small but measurable dip in the star's brightness, known as a transit. These periodic dips are recorded as **light curves**, which form the primary data source for exoplanet detection. However, identifying genuine transit signals is challenging due to stellar variability, instrumental noise, and the massive volume of time-series data. This makes the problem well-suited for data-driven and automated approaches such as machine learning.

1.2 Applications

Exoplanet detection using Kepler data and machine learning has several important applications:

- **Astronomical research**: Improves understanding of planetary formation, evolution, and diversity of planetary systems.
- **Search for habitable worlds**: Helps identify Earth-sized planets in habitable zones, guiding future observational missions.
- **Data-driven astronomy**: Demonstrates the use of machine learning and statistical methods in analyzing large-scale astrophysical datasets.
- **Automation of scientific workflows**: Reduces reliance on manual or rule-based detection pipelines, enabling faster and more scalable analysis.
- **Educational and academic research**: Provides a practical case study for applying machine learning techniques to real-world scientific data.

1.3 Motivation and Need

The primary motivation for this project arises from the limitations of traditional exoplanet detection pipelines. Classical algorithms and rule-based statistical methods, while effective, struggle with noisy data, rare events, and scalability issues when applied to large datasets like those produced by Kepler. Additionally, false positives caused by stellar activity or instrumental artifacts remain a persistent challenge.

Machine learning offers a promising alternative by learning patterns directly from data and adapting to complex, non-linear relationships. Studying the feasibility of such approaches at an early stage helps establish whether automated detection can improve reliability and efficiency. This project is motivated by the need to understand the data, identify challenges, and justify the application of machine learning before proceeding to full-scale implementation.

CHAPTER 02

LITERATURE SURVEY

CHAPTER 02

LITERATURE SURVEY

2.1 Literature Review

This section reviews key research papers relevant to exoplanet detection using Kepler data and machine learning techniques.

1. **Borucki et al. (2010) –** ***Kepler Planet-Detection Mission: Introduction and First Results***
   This paper introduces the Kepler Space Telescope mission and explains the scientific motivation behind using transit photometry for exoplanet detection. The authors demonstrate the mission's early success in identifying Earth-sized exoplanets using high-precision stellar light curves. The study also highlights major challenges such as stellar variability, instrumental noise, and false positives, thereby motivating the need for automated and robust detection techniques.

2. **Jenkins et al. (2010) –** ***Overview of the Kepler Science Processing Pipeline***
   This work describes the end-to-end Kepler data processing pipeline, with particular emphasis on the Transit Planet Search (TPS) algorithm. The paper explains how periodic transit-like signals are detected in noisy photometric time-series data using statistical and rule-based methods. While effective, the study points out limitations related to scalability and sensitivity, especially when dealing with large datasets and weak signals.

3. **Thompson et al. (2018) –** ***Planetary Candidates Observed by Kepler***
   This paper presents a comprehensive catalog of planetary candidates identified from Kepler data. It details the classification, validation procedures, and reliability metrics used to confirm exoplanet candidates. Importantly, the work provides well-labeled datasets that are highly suitable for supervised machine learning applications and for studying biases and limitations in existing detection approaches.

4. **Shallue & Vanderburg (2018) –** ***Identifying Exoplanets with Deep Learning***
   This paper demonstrates the application of deep learning, specifically convolutional neural networks (CNNs), for identifying exoplanet transits in Kepler light curve data. The study shows that neural networks can significantly reduce false positives compared to traditional pipelines by learning complex

transit patterns directly from data. This work establishes the effectiveness of modern machine learning techniques for large-scale exoplanet detection.

5. **Armstrong et al. (2018) – *Exoplanet Detection in the Kepler Dataset Using Machine Learning***
   This study explores the use of supervised machine learning methods for detecting exoplanet signals within the Kepler dataset. The authors analyze feature-based classification approaches and discuss challenges such as class imbalance and noisy signals. The paper highlights the importance of careful feature selection and data preprocessing, which is highly relevant to the objectives of this project.

Together, these works establish Kepler as a reliable data source and clearly justify the exploration of machine learning–based approaches for automating exoplanet transit detection.

2.2 Literature Gap

Despite extensive research, several gaps remain in existing literature:

- Many traditional detection pipelines rely on fixed thresholds and handcrafted features, limiting adaptability to diverse signal patterns.
- While machine learning has been explored, there is a lack of introductory-level studies focusing on feasibility, data understanding, and preprocessing challenges rather than performance optimization.
- Limited emphasis has been placed on clearly documenting the transition from classical statistical methods to data-driven approaches in an educational or academic project context.

This project addresses these gaps by focusing on problem formulation, data characteristics, and conceptual workflow design rather than immediate implementation.

2.3 Objectives of the Project

The main objectives of this project are:

1. To study the fundamentals of exoplanet detection using the transit photometry method.
2. To understand the structure and challenges of Kepler light curve data, including noise and variability.
3. To explore suitable preprocessing techniques such as normalization and detrending.
4. To investigate the feasibility of supervised machine learning approaches for transit detection.

5. To analyze existing literature in order to identify limitations and research gaps.
6. To define a clear methodological framework for subsequent project levels.

2.4 Methodology

The methodology outlines the structured approach proposed for implementing machine learning–based exoplanet detection using Kepler light curve data. This methodology is conceptual and preparatory in nature, aligning with the objectives of Level-01 and Level-02 of the project.

2.4.1 Data Acquisition

Light curve data will be obtained from NASA's Kepler mission archives. The dataset will include stars with confirmed exoplanet transits as well as non-planetary stellar observations. Publicly available and well-documented Kepler datasets ensure reliability and reproducibility of the study.

2.4.2 Light Curve Preprocessing

Raw Kepler light curves often contain noise, systematics, and discontinuities caused by instrumental effects and stellar variability. Preprocessing steps will include normalization of flux values, detrending to remove long-term variations, and handling of missing or corrupted data points. These steps are essential for enhancing transit signal visibility.

2.4.3 Feature Extraction

Relevant features will be extracted from the preprocessed light curves to represent transit characteristics effectively. These may include transit depth, duration, periodicity, and statistical measures such as mean, variance, and skewness. Feature extraction plays a critical role in enabling machine learning models to distinguish between planetary and non-planetary signals.

2.4.4 Machine Learning Model Selection

Supervised machine learning algorithms such as Logistic Regression and Support Vector Machines will be explored as baseline classifiers. These models are chosen for their interpretability and suitability for binary classification problems. The study will focus on understanding model behavior rather than performance optimization at this stage.

2.4.5 Evaluation Strategy

Model performance will be evaluated using appropriate metrics such as accuracy, precision, recall, and confusion matrices. Special attention will be

given to false positives and false negatives, as misclassification can significantly impact exoplanet detection reliability.

CHAPTER 03

CONCLUSIONS AND

LEVEL-2 PLAN OF WORK

CHAPTER 03

3.1 CONCLUSION

This preliminary level of the project establishes a strong conceptual foundation for exoplanet detection using machine learning and Kepler data. Through a detailed study of the problem domain and existing literature, the relevance and feasibility of the chosen topic have been clearly justified.

The survey highlights both the success and limitations of traditional detection methods, reinforcing the need for automated, data-driven approaches. While this level does not focus on implementation or performance evaluation, it provides essential insights into data characteristics, challenges, and methodological considerations. These findings form a solid basis for future work involving detailed data analysis, model development, and experimental validation.

3.2 LEVEL - 2 PLAN OF WORK

The Level-2 phase of this project focuses on moving from conceptual understanding to preliminary implementation and analysis of machine learning–based exoplanet detection using Kepler data. The proposed work plan includes the following stages:

1. **Problem Definition and Requirement Analysis**: Clearly define the exoplanet detection problem as a supervised classification task. Identify project objectives, data requirements, assumptions, and constraints related to Kepler light curve data.

2. **Dataset Selection and Understanding**: Select appropriate Kepler light curve datasets, including confirmed exoplanet candidates and non-planetary signals. Study data structure, metadata, noise characteristics, missing values, and class imbalance.

3. **Data Preprocessing and Conditioning**: Perform preprocessing steps such as normalization, detrending, noise reduction, and handling data gaps. Prepare clean and consistent datasets suitable for machine learning analysis.

4. **Feature Extraction and Representation**: Identify and extract relevant features from light curves, such as transit depth, duration, periodicity, and statistical properties. Explore suitable representations for effective classification.

5. **Preliminary Model Development**: Implement baseline supervised machine learning models such as Logistic Regression or Support Vector Machines to classify exoplanet candidates.

6. **Evaluation and Feasibility Analysis**: Analyze model outputs using appropriate evaluation metrics to assess feasibility and limitations. Study sources of misclassification and false positives.

7. **Model Selection and Refinement**: Compare different models and feature sets to select the most suitable approach for further improvement in the next level.

8. **Documentation**: Prepare a detailed Level-2 report documenting dataset selection, preprocessing steps, model design, results, challenges, and observations to serve as a foundation for Level-3 development.

REFERENCES

1. Borucki, W. J., et al. (2010). *Kepler Planet-Detection Mission: Introduction and First Results*. Science.

2. Jenkins, J. M., et al. (2010). *Overview of the Kepler Science Processing Pipeline*. Astrophysical Journal Letters.

3. Thompson, S. E., et al. (2018). *Planetary Candidates Observed by Kepler*. Astronomical Journal.

4. Shallue, C. J., & Vanderburg, A. (2018). *Identifying Exoplanets with Deep Learning*. Astronomical Journal.

5. Armstrong, D. J., et al. (2018). *Exoplanet Detection in the Kepler Dataset Using Machine Learning*. Monthly Notices of the Royal Astronomical Society.

Blog / Articles:

6. KDAG IIT KGP. *Unveiling the Cosmos: Discovering Exoplanets with Machine Learning*. Medium.