National College of Ireland

Ensemble Machine learning to Detect Exoplanets

MSc Research Project
Data Analytics

Vishal Petkar
Student ID: X21216461

School of Computing
National College of Ireland

Supervisor: Cristina Hava Muntean

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Vishal Petkar
Student ID:	X21216461
Programme:	Masters in Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Cristina Hava Muntean
Submission Due Date:	25/04/2024
Project Title:	Ensemble Machine learning to Detect Exoplanets
Word Count:	6799
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Vishal Petkar
Date:	27th May 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Ensemble Machine learning to Detect Exoplanets

Vishal Petkar
x21216461

Abstract

In last 6 decades the exploration of space has unveiled some of the most profound mysteries of our universe, revealing multitudes of celestial objects and wonders, which includes distant exoplanetary systems. Leveraging on the vast data that was obtained by the earlier Kepler and K2 missions, this report aims to present an approach to detect exoplanets by analysing the fluctuations of light curves with a combination of data from the Kepler, K2 and TESS missions. Four models will be trained – Convolutional Neural Network (CNN) utilizing GPU acceleration, Support Vector Machine (SVM), K-Nearest neighbour and Random Forest to identify the subtle signatures of exoplanetary transits within the fluctuations of light. Post the models achieving a satisfactory performance, an ensemble script combining 3 models was used to evaluate its performance in identifying exoplanets from the light curve data obtained from all 3 sources with test data that the models had never analyzed. The results of this research showcases that an ensemble model with CNN, K-NN and Random forest achieved an accuracy of 0.62, precision of 0.66, recall(sensitivity) of 0.70, specificity of 0.51 and an F1 score of 0.68. This indicates that the ensemble approach, particularly leveraging KNN, exhibits promising performance in accurately identifying exoplanets from the analyzed light curve data, thus contributing significantly to the field of exoplanetary research.

1 Introduction

For centuries, scientists and the common masses have been fascinated over the vast expanse of space. One of the aspects of space that has captured everyone's curiosity is the thought of if there were other planets present outside our solar system. Interestingly, the first ever exoplanet (2 in this instance, named Poltergeist and Phobetor) that was officially discovered in recorded history was just discovered 32 years ago in 1992, and were found orbiting a pulsar. Wolszczan and Frail (1992) The first exoplanet that was discovered orbiting a solar-like star was discovered in 1995 named 51 Pegasis b, which was a massive planet and at the time due to the limitation of technology, only such massive planets could be detected. Queloz and Alsari (2020)

There has been a drastic improvement in technology in the last 3 decades. Both ground based and space based telescope like Kepler and Hubble have contributed greatly in the discovery of exoplanets and in the overall understanding of the universe. The Kepler mission was designed to be a statistical mission to find as many earth sized planets outside our solar system, which are present near or within a stars habitable zone. Malik et al. (2021) In its complete mission's lifespan, the Kepler and K2 were able to observe approximately 530,506 stars continuously for several years. The mission finally ended

in 2018 when the spacecraft ran out of fuel. Following in the success of the Kepler/K2 missions, NASA had launched another mission named the Transiting Exoplanet Survey Satellite (TESS) in 2018. Unlike its predecessor Kepler, TESS was able to cover a much wider area of the sky in its observation and was able to observe a much broader range of stars. Its main mission objective is to survey the closest and brightest stars around earth for transiting exoplanets. The mission was designed to last for 2 years, but continues to operate even now. By mid-November 2023, it has managed to discover 6977 exoplanet candidates of which 402 are confirmed as exoplanets. [1]

The method employed in this research paper to find exoplanets, would be by using the Transit method. The term 'transit' in astronomy context can be defined as the event when a planet passes in front of a star, which is being observed from earth. The Figure 1 shows how a transit light curve occurs in space. When this event happens, the light of the star appears to dim. By measuring this dip in the brightness of a stars luminosity, scientists can determine if the dip in brightness is due to an orbiting planet or due to stellar debris.

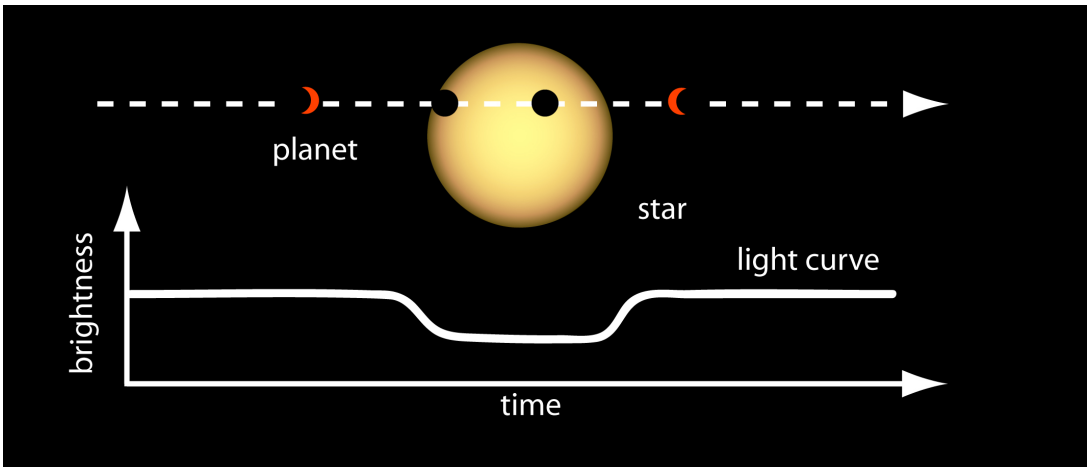

Figure 1: Light curve of a planet transiting its Star.[2]

This can be determined by plotting the light curve of the star. When the brightness of the star dips below a certain threshold indicating a transit event, it is termed as a Threshold Crossing Event (TCE). A threshold is often determined based on statistical considerations, such as signal to noise ratio (SNR), which means that if there is a high SNR threshold then only light curves with more significant dips in brightness are considered as potential transits. McCauliff et al. (2014) For example, an astronomers might set a threshold value such that any decrease in brightness exceeding, 3 to 5 times the standard deviation of the noise in the light curve is considered a potential transit event. This research project holds vital importance within the field of exoplanetary science, by building upon centuries of scientific curiosity and technological advancements. The historical context provided above underscores the transformative impact of recent decades of developments, particularly in the detection of exoplanets. Leveraging the cutting-edge machine learning and data analysis techniques, this research aims to significantly enhance our capacity to detect exoplanets accurately and efficiently. By contributing to the ongoing exploration of space, the discovery of exoplanets continues to broaden our

[1] https://exoplanets.nasa.gov/tess/

[2] Source: `https://exoplanets.nasa.gov/resources/280/light-curve-of-a-planet-transiting-its-star/`

understanding of the universe and the prevalence of planetary systems beyond our own. The methodology employed, particularly the use of the transit method, which represents a powerful tool for identifying exoplanets and probing their fundamental properties. Through meticulous analysis and validation, this study not only advances our understanding of exoplanetary systems but also refines and optimizes detection techniques for future missions.

The process of vetting the light curves of each of the stars observable from earth is a long and tedious one. To tackle this, it is necessary to work with multiple sources of data and make use of multiple machine learning methods that can systematically analyse the data, remove any misleading sources by using pre-set parameters such as TCE values, generate the light curves and is able to classify if the generated light curve is in-fact representative of the presence of an exoplanet or not.

From the above given research problem, the main objectives of this research paper is to answer the below questions.

1. *How much does the combined utilization of Convolutional Neural Network (CNN), Support Vector Machine (SVM), k-Nearest Neighbors, and Random Forest models provide an improvement in exoplanetary signature detection in terms of evaluation metrics such as accuracy, specificity, recall and F1 Scores?*

2. *To what extent does the integration and simultaneous utilization of multiple datasets from Kepler, K2, and TESS contribute to the enhancement in accuracy and reliability of exoplanetary signature detection and classification, as evidenced by comparative performance metrics against data obtained from individual sources?*

This research paper is organized as follows: Section 2 provides a review of related work and contributions in this domain. Section 3 introduces the research methodology and explains the steps taken during exploratory data analysis. Section 4 defines the implemntation of the different machine learning models that were utilized in this research. The evaluation of the ensemble models are presented in Section 5. Section 6 discusses on the results obtained from the different case studies. And finally, Section 7 concludes the research and discusses on possible future works.

2 Related Work

In this section, the various research approaches in using machine learning to detect exoplanets will be discussed

2.1 Machine learning advancements in Astronomical data-analysis

Salinas et al. (2023) In their approach introduced a deep learning architecture that was specifically designed to analyse light curves from astronomical data. It utilized a self-attention mechanism that was inspired by Transformer models which enabled to capture significant features and provide interpretability. In their research they addressed the limitations of traditional CNNs, and put forth their deep learning approach which had simplified manual examination of transit light curves and had achieved competitive results which in turn offered a promising direction for further research in exoplanet science. In all their experiments involving SVM and multilayer perceptron, both models did not

exceed a precision threshold over 0.6. Although their approach was promising, it still had drawbacks. Their approach suffered from scalability issues in the analysis of longer light curves and suffered from potential challenges in generalizing other classifications of astronomical events. In contrast to this, an alternative research was done Schanche et al. (2019)which had focused on ground based wide field transit survey data. Particularly the data from the Wide Angle Search for Planets (WASP) program. They employed a combination of ML methods that included Random Forest Classifiers (RFCs) and CNNs to automate signal classification. Their research achieved an accuracy of approximately 90% in the identification of exoplanets, thereby emphasizing the efficiency and scalability of Machine Learning architectures in handling huge datasets. It also aided in the rapid classification and in reducing manual analysis labour on human observers. While they had achieved high accuracy rates, their approach had still exhibited a very notable fraction of false positives. Additionally their reliance on just using the WASP dataset may limit the quality of classification based on the data available. Both research papers highlight the importance of machine learning techniques in tackling challenges such as the large volume and complexity of astronomical data.

2.2 Exoplanet Detection: Kepler

In a similar way, Cuéllar et al. (2022) the research focused on deep learning techniques that particularly employed convolution neural networks (CNNs), that enhance transit detection within the Kepler Telescope light curves. The model they made was trained on a combination of real and synthetic data which made use of 2D phase folding which was used for feature extraction. Their approach had achieved a superior performance when compared to other existing methods, thereby emphasizing the usefulness of incorporating synthetic data. Their proposed model achieved an accuracy of 0.98, precision of 0.97 and recall of 0.99 in their test. The synthetic data had improved the model's knowledge and performance. However, their approach also consisted within them some drawbacks, such as being limited to single-transit detection. They also required their results to be validated on datasets beyond the Kepler data. In contrast, a research study was done Malik et al. (2021) that introduced a novel machine learning based approach that diverged from the standard deep learning methodologies. Their study utilized the TSFRESH time series analysis library and a gradient boosting classifier. It showcased a competitive performance when it came to detecting exoplanets and had particularly demonstrated a reduction in predicting false positives when compared to conventional algorithms. Their approach also offered an increase in computational efficiency without the need for specialized software. While their approach does offer computational benefits, it still needs the assistance of human supervision and validation on the unseen data. Both these studies represent a significant advancement in the detection of exoplanets, the former highlights the scalability and accuracy achieved from deep learning techniques, while the later showcased the lightweight nature and practical applicability of ML methods. At the same time, both these methods suffer challenges in generalising to data from a broader dataset and requires more refinement with the validation of their results.

More research was done on exoplanet detection and validation. In one study the researchers focused on the development of techniques for the identification of background false positives in Kepler data. They approached this problem by utilizing centroid analysis, PRF-fit technique and photometric centroid technique. Bryson et al. (2013) PRF-fit

stands for Point Spread function fit and is a technique that is used in astronomy to model and analyse the shape and intensity of light sources in an image. Although the mentioned methods make significant contribution to the reliability of the Kepler exoplanet candidate list, they still exhibit certain limitations such as breakdown in low signal to noise scenarios, its sole reliance on the photometric data obtained from Kepler which potentially restricts its application in other datasets. Alternatively, the research done by Armstrong et al. (2021) proposes a novel approach of planet validation through the use of machine learning. Particularly the researchers used a method known as Gaussian process classifier (GPC) as an alternative to the more popular VESPA algorithm, which showed promising results in differentiating between confirmed planets from false positives in the Kepler TCE catalogue, thereby providing rapid validation of thousands of unseen planet candidates. Specifically, the study highlights the precision achieved by machine learning models such as random forest classifiers (RFC), Gaussian process classifiers (GPC), extreme tree classifiers (ET), and multilayer perceptrons (MLP), with AUC metrics ranging from 0.998 to 0.999. However the researchers also identified discrepancies with VESPA algorithm when their model was applied on several candidates, thereby raising concern about relying solely on a single validation method. This limitation cautions against the use of a single validation method which may introduce biases into the model which would lead to misclassification. Therefore, continued research and validation against known datasets is essential to evaluate the reliability and robustness of proposed approach.

2.3 ML with TESS Data

A research study was done which discussed on the potential of the Transiting Exoplanet Survey Satellite (TESS), in its capability to observe the solar system objects and exploring the implications of studying minor planets. Pál et al. (2018) Their study highlighted the capabilities of TESS in providing timeseries imaging data and compares its optical setup with that of the Kepler/K2 mission. Thus emphasizing the larger net expanse of TESS and its differences with Kepler/K2 in data acquisition principles. Their study presented statistics for minor planet transits which affect target star light curves, demonstrating the impact of ecliptic latitudes on the number of encounters. Despite the promising photometry achievable for thousands of minor planets, their research paper acknowledges limitations in detecting fainter objects and the potential confusing effects of minor planet transits on stellar photometry.

Further research was done in the utilization of machine learning methodologies for the detection of exoplanets within NASA's TESS dataset. One study was done using a specific convolutional neural network (CNN) which was named Astronet-Triage-v2. Tey et al. (2023) It was designed to distinguish between eclipsing candidates and other phenomena's within the TESS Full-frame Image (FFI) light curves that were obtained from the TESS data. This network was trained on high-quality data that was curated from the Primary mission and the 1st extended mission of TESS. The Astronet-triage-v2 had exhibited remarkable performance, achieving a recall of 99.6% for transiting events with a precision of 75.7%. It notably outperforms its predecessor, Astronet-Triage. In contrast, a research study was done which introduced a novel AI technique that was developed by ThetaRay, Inc., which combined multiple algorithms that were trained on Kepler data and subsequently validated with confirmed exoplanets before application to TESS data. Ofman et al. (2022) Their research employed the use of semi supervised and unsupervised ML techniques. The ThetaRay system was able to analyse the TESS lightcurves and

was successful in identifying approximately 50 exoplanetary candidates. Although both of these approaches demonstrated a potential for quick identification of exoplanetary candidates, they have some shortcomings. The astronet-triage-v2 lacks the ability to distinguish between transiting exoplanets and eclipsing binaries. Eclipsing binaries are a dual star system where in they orbit each other and periodically eclipse or pass in front of each other from the perspective of earth. This drawback limits the astronet-triage-v2's utility in classification. Conversely, the ThetaRay AI technique need manual validation in order to reduce false positive results. This indicates the need for further optimization and development. Also, its reliance on semi supervised and unsupervised techniques may introduce uncertainties into the model which could cause misclassification.

Another study had focused on classifying exoplanet candidates through transit surveys. Osborn et al. (2020) This was done by leveraging high fidelity simulations that were used to train deep learning models for accurate classification. Their method, while achieving impressive precision of 97.3% and an accuracy of 92% of planets in three-class model, particularly in low signal to noise scenarios, it posed limitations due to their reliance on simulated data which did not fully capture the complexities of real TESS data. This led to necessitating further validation and refinement through training on confirmed TESS planets. On the contrary, the research study done by Vida et al. (2021) tackled the detection of stellar flares in space borne photometric data using recurrent neural networks with LSTM layers. Through training and testing various neural network architectures their study finds that RNNs with LSTM layers perform the best, achieving both high recall and precision rates slightly greater than 70% in detecting flares. However the researchers acknowledge there were several limitations and failures encountered during the experimentation process. Challenges such as model selection, data standardization and convergence of network architecture. Additionally the use of artificial data for training neural networks does raise some concerns about generalizability to real observational data, highlighting the need for further validation on independent datasets, particularly from TESS observations.

Transit timing variations (TTVs) is another way to determine the transit of an exoplanet. In this method the timing of a planets transit across its host star's disk varies over time due to gravitational influences of other bodies in the systems such as other exoplanets or moons. A study was done in detecting these TTVs in the Kepler field by leveraging the observations made by TESS. Jontof-Hutter et al. (2022) Despite the successful recovery of transits from multiple systems and the identification of non-transiting perturbers, their study heavily relies on Kepler data for dynamical constraints, which may limit the robustness of their findings, particularly given the lower signal-to-noise ratio of the TESS transits. Additionally, while pixel-level decorrelation (PLD) enhanced transit detection in noisy TESS data, they faced limitations in detecting transits of faint stars and shallow transits posed challenges, potentially leading to missed detections or lack of validation for certain planets. In contrast another research study was performed where the researchers had evaluated the potential of TESS to detect and characterize planetary systems in the Kepler field. Christ et al. (2019) By modelling the expected transits of confirmed and candidate planets detected by Kepler, the research forecasts TESS's ability to detect these planets and improve our understanding of the planets properties. Their research predicts that TESS has a high probability of detecting a significant number of planets, particularly hot Jupiters, and suggests it as a powerful tool for characterizing transit timing variations (TTVs). However, the study relies on assumptions about TESS's noise properties and contamination ratios, which could affect the accuracy of the predictions.

Additionally, while TESS is expected to enhance measurements of planetary parameters and reduce transit timing uncertainties, there are limitations in detecting multiplanet systems and tidal orbital decay. The paper discusses strategies for maximizing TESS's scientific yield, including extended mission plans, but acknowledges the need for further validation and understanding of TESS's performance in the Kepler field. Overall, while the paper provides valuable insights into TESS's potential contributions to exoplanet research, its reliance on assumptions and the complexity of characterizing multiplanet systems and tidal effects highlight the challenges and uncertainties in predicting TESS's performance accurately.

2.4 Takeaways from Related work

There has been a lot of work done by several researchers in the field of astronomy and in the search of exoplanets. These works range from working upon individual sources of data such as Kepler, Hubble, JWST etc. to different methods of analysis and processing of this available data. Various research teams have used many algorithms in their pursuit of searching and cataloguing new exoplanets. However in most of these study's the researchers are attempting to compare the performance of different algorithms in achieving the same tasks. Also, in most of the studies, only individual sources of data worked upon. Therefore to improve on these two factors, this research study aims to find and showcase the advantage of using a voting system with a combination of 3 different machine learning algorithms and make use of a combined source of data.

3 Research Methodology

The related works section discussed the different data sources/datasets and different algorithms which were used by various researchers in their study to discover exoplanets through the process of transit light curves. Machine learning algorithms such as Astronet-Triage-v2, CNN, and even AI such as ThetaRay were used individually in the analysis of light curve data of stars. In this paper, an ensemble approach is proposed which combines the outputs of 3 machine learning algorithms and uses a voting system to determine the final output. The proposed algorithms to be used in this research are CNN, Random Forest, KNN and SVM. The proposed methodology aims to provide a robust model that can identify and classify exoplanetary transits by analysing a star's light curve. The following sections will discuss on the step by step process of the model creation starting from data selection as shown in the below Figure 2

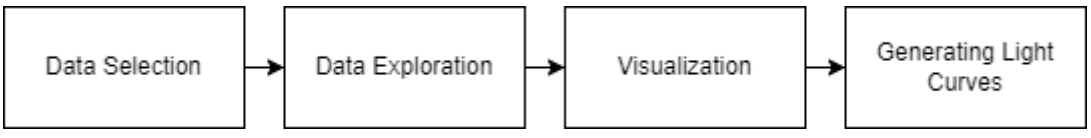

Figure 2: Stages of research methodology

3.1 Data Selection

The NASA Exoplanet Archive was one of the main sources of data used in this research. The Kepler data and TESS data was obtained in the form of CSV files. The data

downloaded was divided into 2 bundles, one containing data of confirmed exoplanets, and the other comprising of confirmed false positives. The CSV file contains several columns of data such as the astronomical names given to stars and planets, the number of stars and planets in the given system, distance of the system from earth and other physical data properties of those stars, such as temperature etc are provided.

3.2 Data Exploration

The csv data file is systematically analysed and exploratory data analysis is performed on it. The "TIC" or "KIC" ID values are initially checked for duplicates and removed. Then the values of specific columns of interest are checked for NA or NULL values. The specific columns that were checked are as follows.

1. sy_snum = Number of Stars
2. sy_pnum = Number of Planets
3. discoverymethod = Discovery Method
4. disc_year = Discovery Year
5. pl_orbper = Orbital Period [days]
6. sy_dist = Distance [pc] (converted to lightyears)
7. st_teff = Stellar Effective Temperature [K]

Once the data has been cleaned of any null values, visualizations were created to better understand and derive knowledge from the data.

3.3 Data Visualization

The Figure 3 below shows that the majority of the exoplanets discovered in the Kepler dataset were discovered by the transit method. A total of 1888 exoplanets that were catalogued were discovered with this approach and remains to be the most widely used method today.

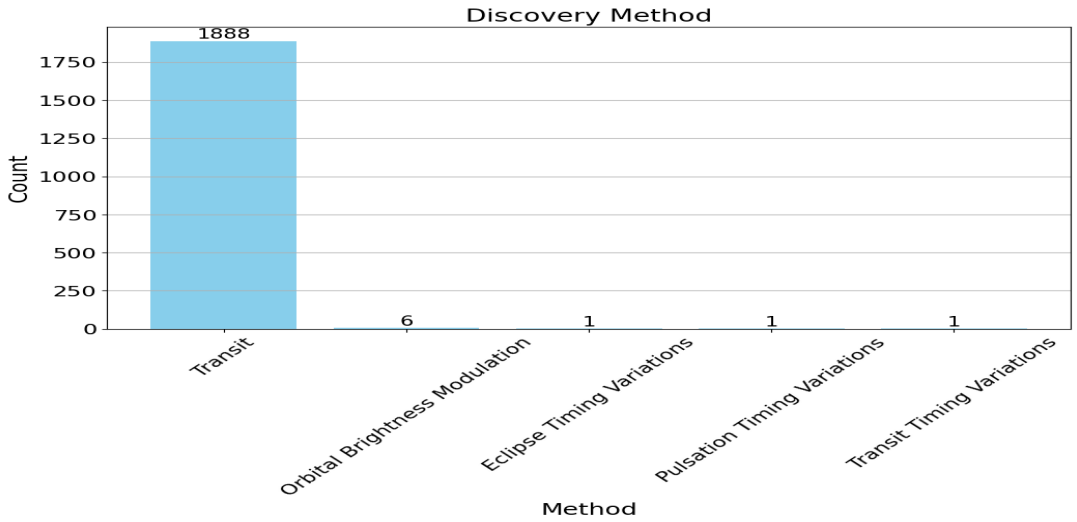

Figure 3: Bar plot of discovery methods used to discover planets

In Figure 4, it can be observed that the most number of exoplanets that were discovered were in the year 2016, with a total of 1141 which accounts for 60.15%, followed closely by 2014 with 16.55%. The steep rise in discoveries in 2016 could be attributed to the advancement in processor chips and in improved algorithmic performances.

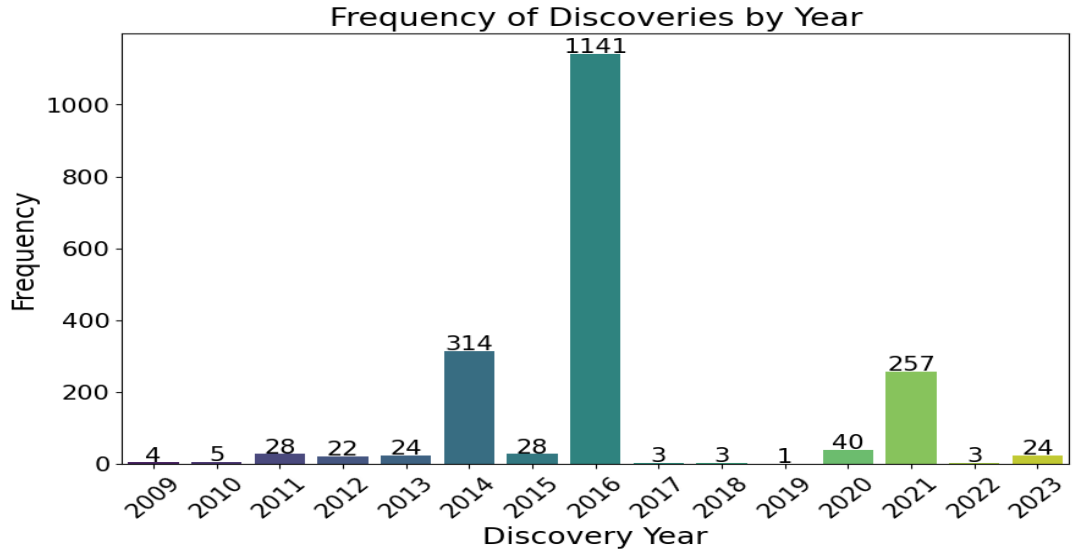

Figure 4: Bar plot of discovery methods used to discover planets

Figure 5 shows that most of the planetary systems that were discovered mainly consisted of just a single planet orbiting its host star. A total of 74.06% of all confirmed exoplanets discovered by Kepler comprises of a single star-planet system.

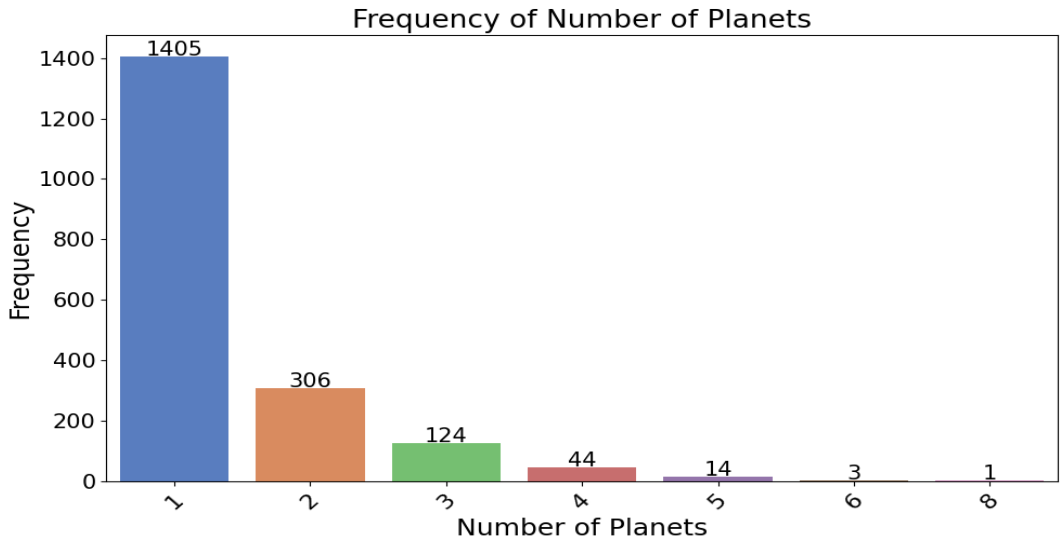

Figure 5: Bar plot of frequency of planets in a system

Figure 6 shows the distance of the exoplanets from earth. Analyzing this data shows that 20.24% of planets lie in the range of 1 to 1500 light years, 61.57% of planets are between the ranges of 1500 and 4000 light years and 18.08% of planets are between 4000 and 10000 light years from earth.

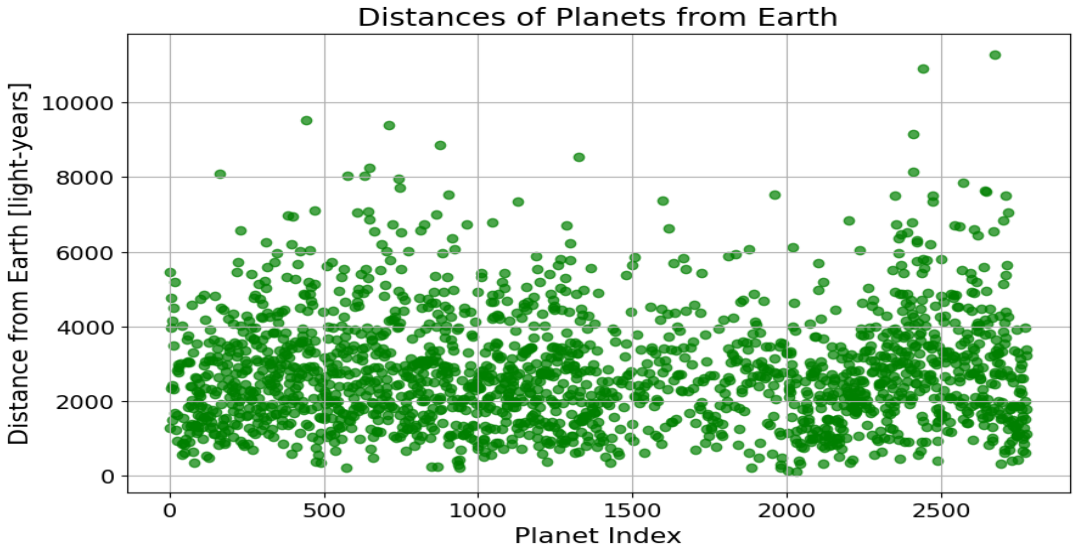

Figure 6: Distance of planets from earth [Light-years]

3.4 Generating Light Curves

Post the initial data exploration, a light curve generating script file was created using Python and the "TIC" and "KIC" ID's of stars were given as input. A dataset containing the various flux values of the given star were then downloaded using the lightkurve python module from the mast website.[3] Once the flux data is available, it can be plotted and

[3]https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html

worked upon as can be seen in the below Figure 7

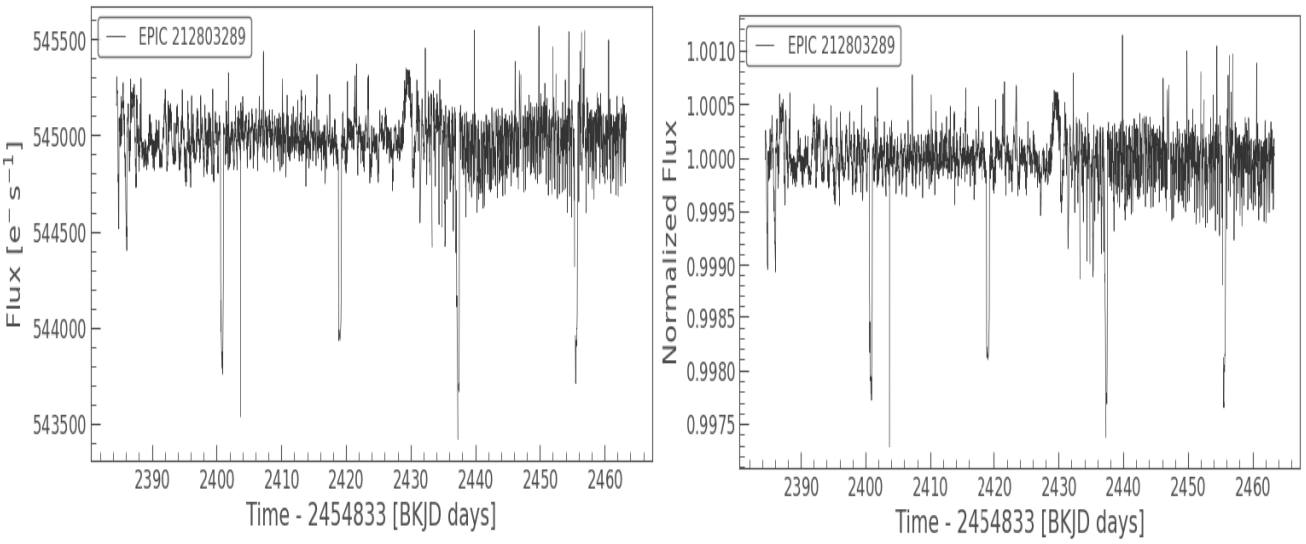

Figure 7: Light curve of a Star's luminosity and Flattened Light curve

The light curve is then flattened using the flatten() command to remove any long term trends. This normalizes the flux values as seen in the Figure 7.

Following this, the to_periodogram() command is used on the flattened lightcurve to check for periodic signals in the time series data. It helps in identifying how the power of a signal is distributed across different frequencies, where the higher peaks indicate periodic signals at those frequencies. In the context of astronomy, periodograms are often used to search for exoplanets by detecting the periodic dimming of a star's light caused by transiting planets. Then the period of max power is calculated which indicates the best fit period of the signal. In the given example, the best fit period is 18.26703 days, which indicates the orbital period of the exoplanet orbiting around this star. Using this information, the signal can then be folded to this "best fit period" or orbital period as seen in Figure 8.

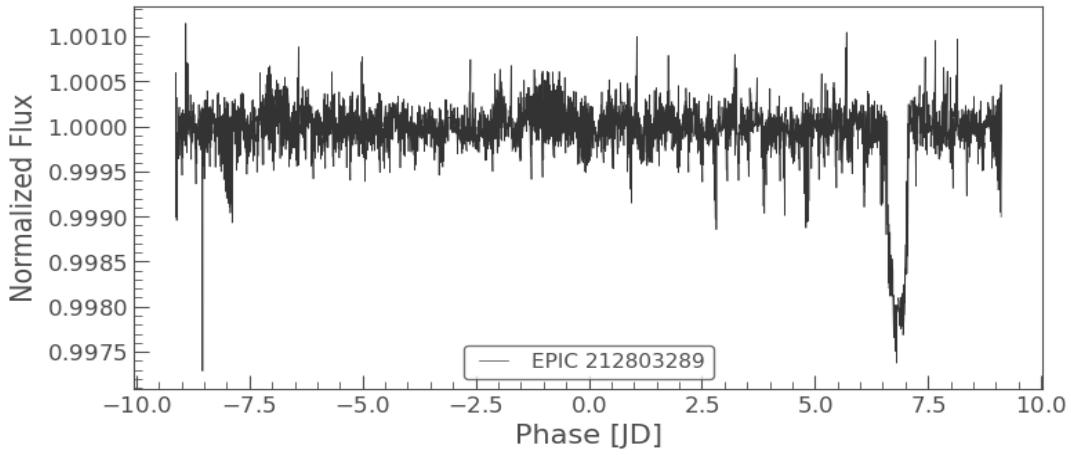

Figure 8: Folded light curve of a Star

The folded light curve shows that there is a significant dip in the flux of star light. To

make the plot appear cleaner, the binning command is used. Binning involves dividing the data into groups or bins and computing the statistics for each bin. Here, the binsize=12 implies that the data will be grouped into bins of size 12 data points each. The summary statistic computed for each bin is typically the mean, median, or sum of the data points within the bin. The binned light curve plot can be seen below in Figure 9.

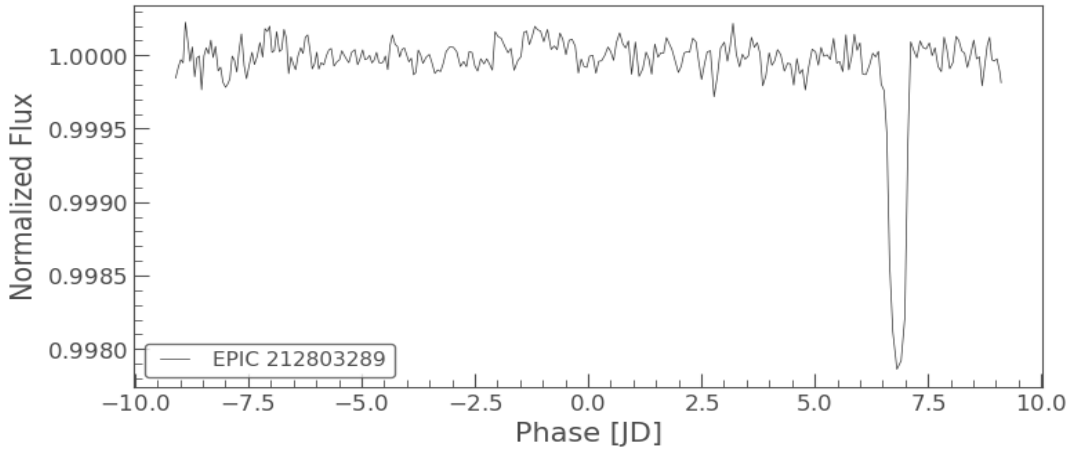

Figure 9: Binned light curve of a Star

It can be observed that upon binning the dip in the light curve flux can be seen more significantly. This graph falls by more than almost 20 units from 1.0000 to slightly below 0.9980. This drop in the flux value is known as Threshold Crossing Event (TCE) McCauliff et al. (2014) In this method the light curves are generated for both csv's of confirmed planets and for confirmed false positives for Kepler, K2 and TESS data. In the confirmed planet list only the star systems that consisted of only a single planet were used to generate light curves. At the end of this process, there were 634 light curve graphs of confirmed planets and 596 light curves of confirmed false positives from Kepler, K2 and TESS data sources. Using this data, the machine learning models were trained.

4 Implementation

Once the light files have been generated, they are split into 2 batches for training and testing, with the train-test ratio's of 70/30 and 80/20. Once the data has been split, they are used to individually train the ML models as shown in the Figure 10 below.

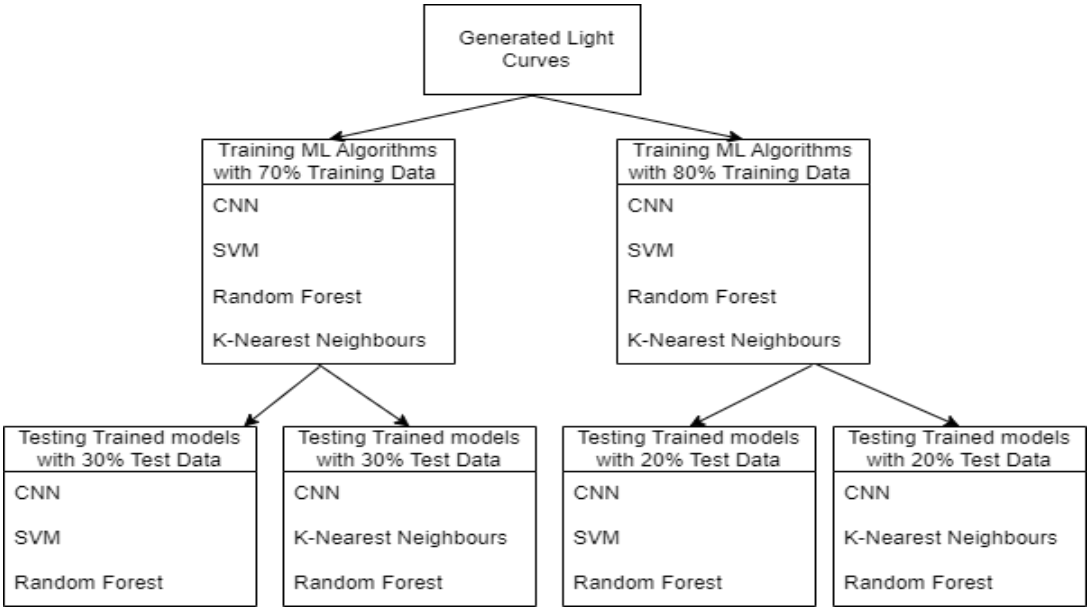

Figure 10: Machine learning models training and testing stages

4.1 Convolutional Neural Networks (CNN)

The first model that was trained was the CNN model. It made use of the Keras API with a TensorFlow backend. The CNN architecture comprises of multiple layers that are designed for image classification tasks. It begins with a convolutional layer with 16 filters, each of size 3x3, employing the ReLU activation function. This layer processes input images of size 256x256 pixels. Following convolution, max-pooling operations are applied to reduce spatial dimensions. The pattern of convolution followed by max-pooling is repeated several times, with varying numbers of filters (32 and 16) in alternate layers, this was aimed at capturing hierarchical features within the image. After the final convolutional layer, a flattening operation is performed to convert the 2D feature maps into a 1D vector. This vector is then fed into fully connected layers consisting of 256 and 128 neurons, each activated by the ReLU function, facilitating feature combination and abstraction. The network concludes with a single neuron output layer activated by the sigmoid function, which is suitable for binary classification tasks. The Adam optimizer is employed for training, using the binary cross-entropy loss function which is used to measure the disparity between predicted and actual classifications. Additionally, the training process is monitored and visualized using TensorFlow's TensorBoard utility, with training and validation data provided for the necessary epochs. The model is trained using the onboard available system Nvidia GPU.

4.2 Random Forest Classifier

The 2nd model to be trained was the Random forest classifier which was designed to be a supervised machine learning tasks, particularly for classification. In this model the RF is instantiated with 170 decision trees and is seeded for reproducibility with a random state parameter set to 42. To assess the models performance and generalization ability, k-fold cross-validation is employed where K is set to 20. This ensures comprehensive evaluation across various subsets of the dataset. The 'accuracy' scoring metric was chosen in order to evaluate its performance. During the cross-validation process, the dataset is split into

k-subsets and each subset is iteratively treated as a validation set, while the remaining data is used for training. Performance metrics such as accuracy, precision, recall, and F1-score are computed for each fold, which helps in providing a comprehensive insight into the classifier's behavior across different subsets of the data. This approach ensures robustness and reliability in assessing the classifier's performance, as it considers multiple splits of the data for both training and evaluation. Overall, this methodology approach offers a rigorous evaluation framework for the Random Forest classifier.

4.3 K-Nearest Neighbour (K-NN)

Another model that was trained was the K-nearest neighbour model. It loads the pre-processed image data, converts the images into flattened arrays and their respective class labels into one-hot encoded format. Utilizing scikit-learn's KNeighborsClassifier, the code proceeds to train the k-NN model using k-fold cross-validation, wherein the data is split into 20 folds for training and evaluation. Performance metrics such as accuracy, precision, recall and F1 score are computed for each of the folds and subsequently averaged to gauge the overall models performance.

4.4 Support Vector Machine (SVM)

The final model to be trained is the Support Vector machine (SVM). It was used due to its capability in classification tasks, and due to its ability to employ k-fold cross validation approach to evaluate its performance. Initially a pipeline is constructed, which encapsulates both feature scaling through standardization and the SVM classifier itself. The SVM is configured to utilize a linear kernel and output probability values. Also a 10 fold cross validation strategy was employed, to ensure thorough evaluation of the SVM model's performance while also trying to mitigate potential biases in the assessment. Throughout each iteration of cross-validation, the dataset was split into training and testing subsets, with the SVM model trained on the former and evaluated on the latter. Performance metrics including accuracy, precision, recall, and F1-score are computed for each fold, which provides a comprehensive insight into the classifier's effectiveness across diverse data partitions. This rigorous evaluation methodology ensures the robustness and reliability of the SVM classifier, essential for yielding credible results in research investigations.

4.5 Ensemble Code

Once all the models are created, they are loaded into an ensemble script. The models are then tested against the test data that was initially kept separate from them during the training phase. The models each take the same input file and produce their respective outputs, which are saved in a file. Predictions are made by each model for every image, and a voting mechanism is employed to determine the final classification decision based on the majority vote among the models. Evaluation metrics such as accuracy, precision, recall, F1-score, and specificity are computed using sklearn.metrics functions. Finally, the metrics are printed and visualized through a heatmap of the confusion matrix. This all-inclusive approach enables thorough assessment and comparison of the three models performance, ensuring it's robustness and reliability in the classification task.

5 Evaluation

Evaluation is the pivotal stage in every machine learning endeavor. It facilitates an understanding of the model's performance and verifies it's intended functionality. To evaluate the models performance, there are 4 different experiments/case studies that will be discussed ahead.

5.1 Case Study 1 – 70/30 Train Test split (SVM, CNN, RF)

As discussed in the previous section, the data was split into batches for training and testing. Three models were trained on this 70% training data. Their individual performances are shown below in Table 1.

Table 1: Evaluation Metrics of Different Models

Model Name	Mean Accuracy	Mean Precision	Mean Recall
CNN	0.742	0.733	0.750
Random Forest	0.599	0.613	0.599
SVM	0.581	0.579	0.581

The CNN model shows the most promising results among the 3 models. The loss function of the CNN model falls as expected, and conversely the accuracy rose during its training. As can be observed, the loss function has fallen from 0.70 to less than 0.55. In the same way the accuracy function has risen from 0.5 to 0.7 during the CNN model's training, as observed in Figure 11.

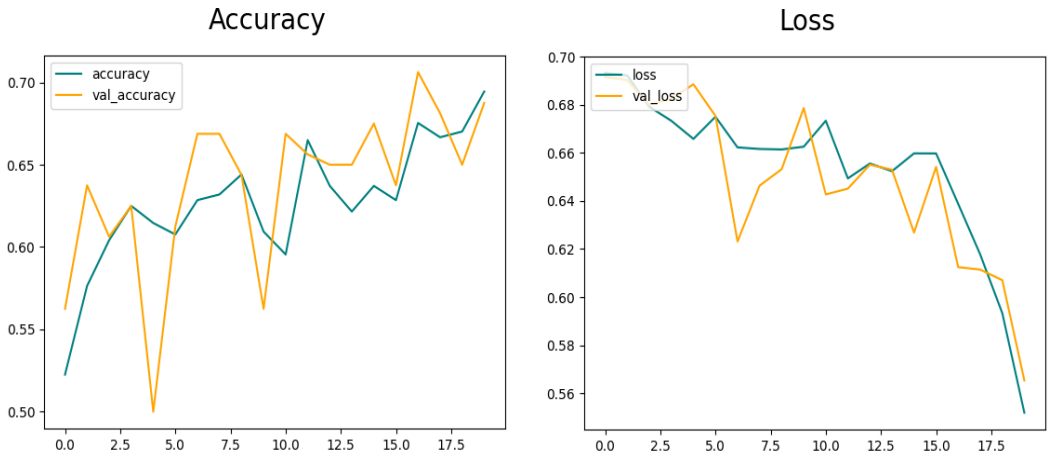

Figure 11: Accuracy and Loss of CNN model

The 3 models are then fed the test data to predict the following confusion matrix in Figure 12

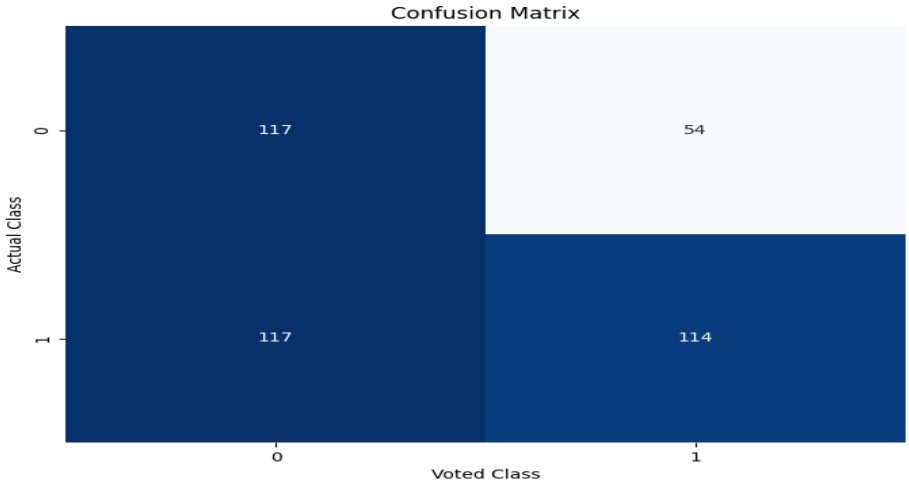

Figure 12: Confusion Matrix of Ensemble model (CNN, SVM, Random forest)

As can be observed from the above confusion matrix, the ensemble of the 3 models have successfully classified 117 light curves as an exoplanet (represented as 0) and 114 light curves as a True Negative (false positive exoplanet signal represented as 1). The Table 2 shows the evaluation metrics of the ensemble model.

Table 2: Evaluation Metrics of the Ensemble Model

Metric	Value
Accuracy	0.57
Precision	0.68
Recall (Sensitivity)	0.49
Specificity	0.68
F1 Score	0.57

The ensemble model demonstrates a moderate level of performance as observed by its accuracy of 0.57, which indicates that it was able to correctly predict the outcomes of 57% of instances. Precision, measuring the proportion of correctly identified positive cases among all instances predicted as positive, is at 0.68, suggesting that when the model predicts a positive outcome, it is correct around 68% of the time. However, the model's recall, also known as sensitivity, is comparatively lower at 0.49, signifying that it captures only 49% of all actual positive cases. On the other hand, specificity, representing the proportion of correctly identified negative cases among all instances predicted as negative, mirrors the precision score at 0.68. The F1 score, which balances precision and recall, aligns with the accuracy at 0.57, indicating a harmonious blend of precision and recall but with room for improvement in capturing true positives and minimizing false negatives.

5.2 Case Study 2 – 80/20 Train Test split (SVM, CNN, RF)

In the 2nd case study, the training of the models were done with 80% of the data while keeping 20% hidden away for testing. The evaluation metrics obtained are given in the below Table 3.

Table 3: Mean Evaluation Metrics of Different Models

Model Name	**Mean Accuracy**	**Mean Precision**	**Mean Recall**
CNN	0.681	0.658	0.806
Random Forest	0.600	0.615	0.600
SVM	0.557	0.558	0.557

Again the CNN model shows the best scores in terms of accuracy, precision and recall. The loss function drops as expected during the training process and the accuracy increases as shown in Figure 13

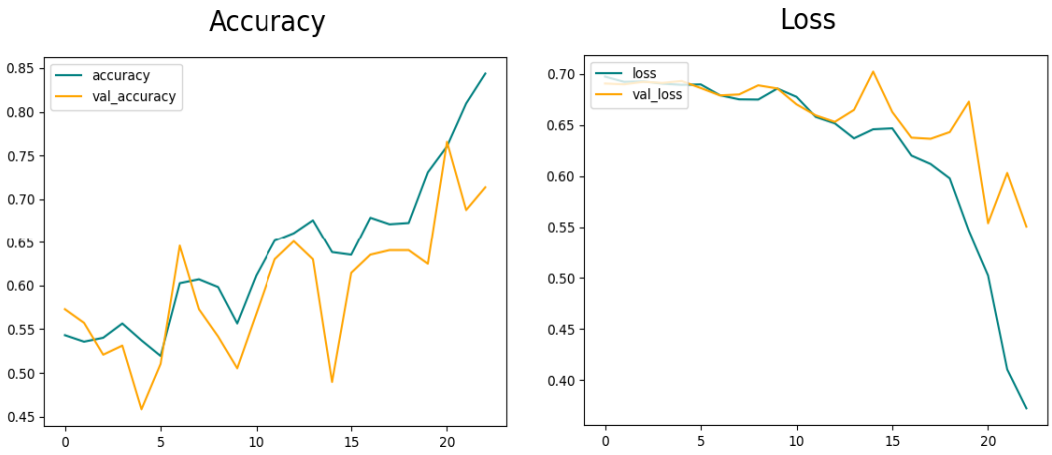

Figure 13: Accuracy and Loss of CNN model

The 3 models are then fed the test data to predict the following confusion matrix in Figure 14

As can be observed from the above confusion matrix in Figure 14, the ensemble of the 3 models have successfully classified 81 light curves as an exoplanet (represented as 0) and 59 light curves as a True Negative (false positive exoplanet signal represented as 1).

Table 4: Evaluation Metrics of the Ensemble Model

Metric	Value
Accuracy	0.57
Precision	0.66
Recall (Sensitivity)	0.44
Specificity	0.73
F1 Score	0.53

The above evaluation results in Table 4 shows that the ensemble model has an accuracy of 0.57 indicates that the model correctly predicted 57% of the instances. Precision, measuring the proportion of correctly predicted positive cases among all instances classified as positive, is reported at 0.66, suggesting a relatively high accuracy in identifying true positives. However, the recall, or sensitivity, stands at 0.44, indicating that the model captured only 44% of all positive instances. Specificity, which reflects the ability to correctly identify negative cases, is reported at 0.73, demonstrating a notable capability to avoid false positives. The F1 score, a harmonic mean of precision and recall,

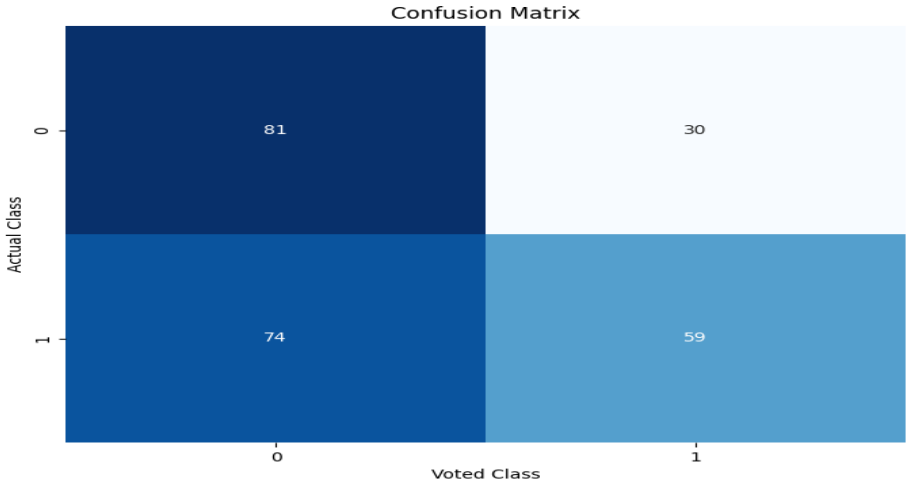

Figure 14: Confusion Matrix of Ensemble model (CNN, SVM, Random forest)

is calculated as 0.53, indicating a balance between precision and recall. Overall, while the model demonstrates respectable precision and specificity, improvements are needed in recall to enhance its ability to capture all positive instances effectively.

5.3 Case Study 3 – 70/30 Train test split (K-NN, CNN, RF)

In case study 3, the K-NN model is used in place if the SVM model. Here all the models are trained using 70% of the complete data. On substituting the models we get the following evaluation metrics for the 3 models.

Table 5: Mean Evaluation Metrics of Different Models (CNN, KNN and Random Forest)

Model Name	**Mean Accuracy**	**Mean Precision**	**Mean Recall**
CNN	0.742	0.733	0.750
Random Forest	0.599	0.612	0.599
K-NN	0.527	0.517	0.527

When these 3 models are given the 30% test data, the resulting confusion matrix is as shown in the below Figure 15

As observed in the above Table 6, with the accuracy value of 0.62, the model is able to correctly classify approximately 62% of the instances indicating the overall correctness of a moderate level. Precision, is reported as 0.66, which measures the proportion of correctly predicting positive instances among all instances predicted as positive. The recall or sensitivity stands at 0.70 indicating that the model was able to identify 70% of all actual positive instances, demonstrating the models capability to capture relevant data points. Specificity, representing the proportion of correctly predicted negative instances among all instances predicted as negative, is noted at 0.51, which suggests a relatively weaker performance in accurately identifying negative instances. Finally, the F1 score, which combines precision and recall into a single metric, is computed as 0.68, indicating a balanced trade-off between precision and recall. This ensemble model exhibits a satisfactory level of performance, with the notable strengths in recall and precision, but with room for improvement in specificity.

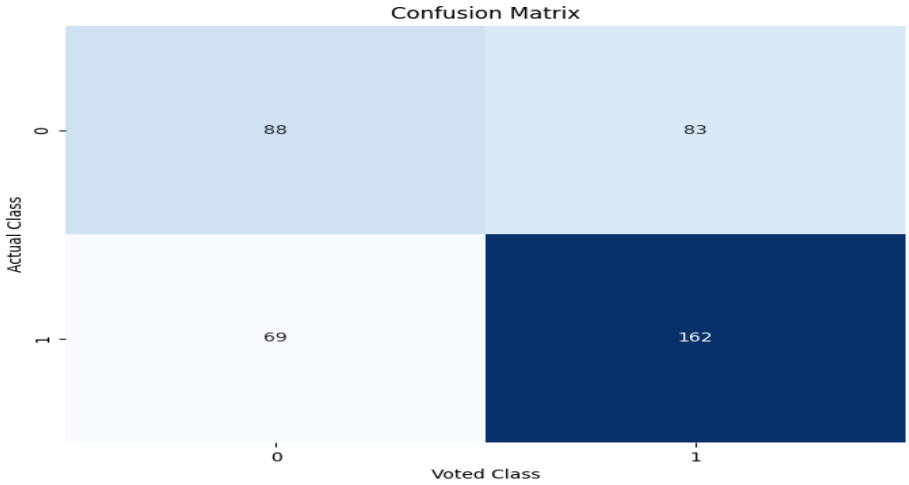

Figure 15: Confusion Matrix of Ensemble model (CNN, K-NN, Random forest)

Table 6: Evaluation Metrics of the Ensemble Model

Metric	Value
Accuracy	0.62
Precision	0.66
Recall (Sensitivity)	0.70
Specificity	0.51
F1 Score	0.68

5.4 Case Study 4 – 80/20 Train test split (K-NN, CNN, RF)

In case study 4, the K-NN model is used in place of the SVM model. These models were trained using 80% of the main data. The evaluation metrics for these 3 models are given below in Table 7.

Table 7: Mean Evaluation Metrics of Different Models

Model Name	**Mean Accuracy**	**Mean Precision**	**Mean Recall**
CNN	0.681	0.658	0.806
Random Forest	0.600	0.615	0.600
K-NN	0.562	0.572	0.562

These 3 models are given the remaining 20% test data which results in the following confusion matrix as shown below in Figure 16.

According to the obtained evaluation metrics in Table 8, the ensemble model is able to provide an accuracy of 0.60, which indicates that approximately 60% of the predictions made by the model were correct. The precision also stands at 60%. The recall or sensitivity is reported at 0.79, demonstrating the model's capability to capture about 79% of all actual positive instances, showing a strong ability to detect relevant data points. However, the specificity of the ensemble model is noted at 0.38, which indicates a relatively weak performance in accurately identifying negative instances. Finally, the F1 score, which balances precision and recall, is computed as 0.68, indicating a reasonable trade-off between these two metrics. Overall, the ensemble model exhibits notable

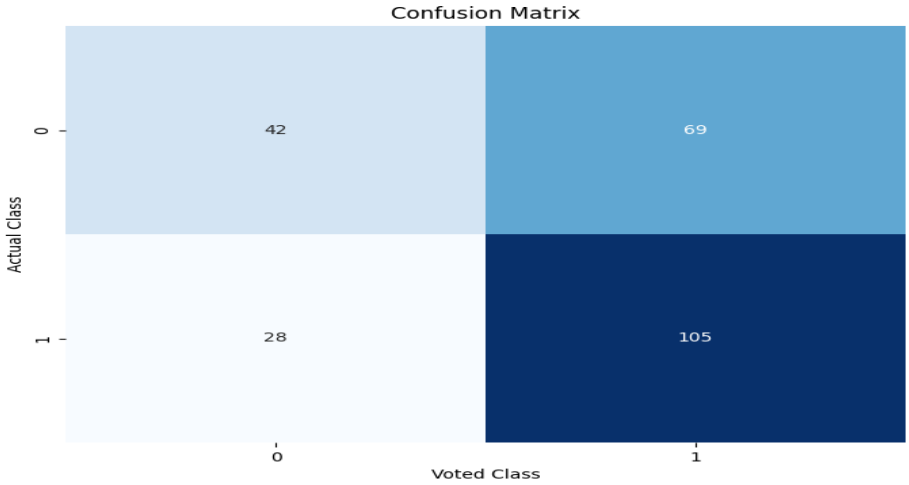

Figure 16: Confusion Matrix of Ensemble model (CNN, K-NN, Random forest)

Table 8: Evaluation Metrics of the Ensemble Model

Metric	Value
Accuracy	0.60
Precision	0.60
Recall (Sensitivity)	0.79
Specificity	0.38
F1 Score	0.68

strengths in recall, but there is room for improvement in specificity to enhance overall performance.

5.5 Discussion

After comparing the evaluation metrics of all 4 case studies, it is evident that the ensemble models each demonstrate strengths and weaknesses across the various performance metrics. The first 2 ensemble models, employing SVM exhibit similar accuracies of 0.57, with a precision value of 0.68 and 0.66 respectively. However these 2 models vary notably in terms of recall and specificity. The first model is able to achieve a recall of 0.49 which is higher than the 2nd model's recall at 0.44. Whereas, the first model is able to achieve a specificity of 0.68 which is lower compared to the 2nd model's specificity at 0.73. In contrast to this, the latter two models utilize K-nearest neighbours instead if SVM, and showcase a higher performance of accuracy of 0.62 and 0.60 respectively, suggesting that their classification performance is better overall compared to the SVM models. Furthermore the KNN models are able to demonstrate superior sensitivity(recall) values, which shows their effectiveness in correctly identifying positive instances (light curve is an exoplanet), with values of 0.70 and 0.79 respectively.

However it is important to also note that the KNN models exhibited lower specificity values, which suggests higher rate of false positives when compared to the SVM based ensemble models. Overall considering the comprehensive evaluation metrics, the KNN-based ensemble models show promising performance, particularly in terms of accuracy and sensitivity, making them preferable for applications where correctly identifying pos-

itive exoplanet instances is crucial.

The case studies were designed in such a way to test the following hypotheses and see if there are any noticeable differences in the final set of models –

1. Would there be any difference in the models evaluation metric values if the data was trained with 70% or 80% of the complete dataset.

2. As CNN and Random forest algorithms have been used in experimentations/research such as this before, would the addition of SVM or K-NN add any improvement in the final results obtained.

3. Between SVM and K-NN algorithm, which algorithm is better suited for classification of lightcurves.

To check on these hypotheses, the experiment was divided into 4 test cases and the final results obtained in each test case was compared.

6 Conclusion and Future Work

There is a lot of potential in improving the effectiveness and accuracy of the models to detect exoplanets and discern them from false positive light signals. According to the findings of this research, the ensemble models that utilize Support Vector Machine (SVM) and K-nearest neighbours, in the search for exoplanets using light curves, reveals intriguing insights. The SVM based models showcase a consistency in accuracy and precision, but show varying levels of recall and specificity. On the other hand, the KNN based models exhibit superior accuracy and sensitivity, indicating their proficiency in correctly identifying positive instances of exoplanets. However, this advantage comes with a trade-off as they demonstrate lower specificity, potentially leading to a higher rate of false positives. Considering the evaluation metrics, the KNN based ensemble models emerge as the more promising model in scenarios where accurately identifying positive exoplanet instances is paramount.

Secondly, the models that were trained with 70% of the data showed an overall better performance than the models trained with 80% of the data. This is true in both cases of ensemble models where SVM and KNN was used. The only exception is that in KNN ensemble based models, the models trained with 80% data showed better recall of 0.79 compared to 0.70. Hence the train test split ratio of 70/30 is the better choice when training models.

Because the data used in this study was from different sources, future work can be done in applying the same ensemble models for a newer dataset such as the JWST dataset. Additionally, other models such as pretrained models, and other classification models can be trained and put together in an ensemble network and its performance and accuracy can be checked against detecting exoplanets in the light curves of distant stars. The main stakeholders that would benefit from this research would be researchers from the Astronomical discipline, space agencies such as NASA or ISRO and other data scientists with an interest in astronomy.

This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and

Space Administration under the Exoplanet Exploration Program.

References

Armstrong, D. J., Gamper, J. and Damoulas, T. (2021). Exoplanet validation with machine learning: 50 new validated kepler planets, *Monthly Notices of the Royal Astronomical Society* **504**(4): 5327–5344.

Bryson, S. T., Jenkins, J. M., Gilliland, R. L., Twicken, J. D., Clarke, B., Rowe, J., Caldwell, D., Batalha, N., Mullally, F., Haas, M. R. et al. (2013). Identification of background false positives from kepler data, *Publications of the Astronomical Society of the Pacific* **125**(930): 889.

Christ, C. N., Montet, B. T. and Fabrycky, D. C. (2019). Observations of the kepler field with tess: Predictions for planet yield and observable features, *The Astronomical Journal* **157**(6): 235.

Cuéllar, S., Granados, P., Fabregas, E., Curé, M., Vargas, H., Dormido-Canto, S. and Farias, G. (2022). Deep learning exoplanets detection by combining real and synthetic data, *Plos one* **17**(5): e0268199.

Jontof-Hutter, D., Dalba, P. A. and Livingston, J. H. (2022). Tess observations of kepler systems with transit timing variations, *The Astronomical Journal* **164**(2): 42.

Malik, A., Moster, B. P. and Obermeier, C. (2021). Exoplanet detection using machine learning, *Monthly Notices of the Royal Astronomical Society* **513**(4): 5505–5516.
URL: *https://doi.org/10.1093/mnras/stab3692*

McCauliff, S., Jenkins, J. M., Catanzarite, J., Burke, C. J., Coughlin, J. L., Twicken, J. D., Tenenbaum, P., Seader, S., Li, J. and Cote, M. (2014). Automatic classification of kepler threshold crossing events, *arXiv preprint arXiv:1408.1496* .

Ofman, L., Averbuch, A., Shliselberg, A., Benaun, I., Segev, D. and Rissman, A. (2022). Automated identification of transiting exoplanet candidates in nasa transiting exoplanets survey satellite (tess) data with machine learning methods, *New Astronomy* **91**: 101693.

Osborn, H. P., Ansdell, M., Ioannou, Y., Sasdelli, M., Angerhausen, D., Caldwell, D., Jenkins, J. M., Räissi, C. and Smith, J. C. (2020). Rapid classification of tess planet candidates with convolutional neural networks, *Astronomy & Astrophysics* **633**: A53.

Pál, A., Molnár, L. and Kiss, C. (2018). Tess in the solar system, *Publications of the Astronomical Society of the Pacific* **130**(993): 114503.

Queloz, D. and Alsari, M. (2020). The discovery of the first exoplanet orbiting a solar-type star, *Scientific Video Protocols* .
URL: *https://doi.org/10.32386/scivpro.000017*

Salinas, H., Pichara, K., Brahm, R., Pérez-Galarce, F. and Mery, D. (2023). Distinguishing a planetary transit from false positives: a transformer-based classification for planetary transit signals, *Monthly Notices of the Royal Astronomical Society* **522**(3): 3201–3216.

Schanche, N., Cameron, A. C., Hébrard, G., Nielsen, L., Triaud, A., Almenara, J., Alsubai, K., Anderson, D., Armstrong, D., Barros, S. et al. (2019). Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys, *Monthly Notices of the Royal Astronomical Society* **483**(4): 5534–5547.

Tey, E., Moldovan, D., Kunimoto, M., Huang, C. X., Shporer, A., Daylan, T., Muthukrishna, D., Vanderburg, A., Dattilo, A., Ricker, G. R. et al. (2023). Identifying exoplanets with deep learning. v. improved light-curve classification for tess full-frame image observations, *The Astronomical Journal* **165**(3): 95.

Vida, K., Bódi, A., Szklenár, T. and Seli, B. (2021). Finding flares in kepler and tess data with recurrent deep neural networks, *Astronomy & Astrophysics* **652**: A107.

Wolszczan, A. and Frail, D. A. (1992). A planetary system around the millisecond pulsar psr1257+ 12, *Nature* **355**(6356): 145–147.