

# Detecting Exoplanets from light curves of Kepler Mission using Fast Fourier Transform and Recurrence Plots

MSc Research Project  
Data Analytics

Aanchal Singh  
Student ID: X19221771

School of Computing  
National College of Ireland

Supervisor: Majid Latifi

National College of Ireland  
Project Submission Sheet  
School of Computing



<b>Student Name:</b>	Aanchal Singh
<b>Student ID:</b>	X19221771
<b>Programme:</b>	Data Analytics
<b>Year:</b>	2021
<b>Module:</b>	MSc Research Project
<b>Supervisor:</b>	Majid Latifi
<b>Submission Due Date:</b>	16/08/2021
<b>Project Title:</b>	Detecting Exoplanets from light curves of Kepler Mission using Fast Fourier Transform and Recurrence Plots
<b>Word Count:</b>	7513
<b>Page Count:</b>	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.

<b>Signature:</b>	Aanchal Singh
<b>Date:</b>	22nd September 2021

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

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

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

<b>Office Use Only</b>	
Signature:	
Date:	
Penalty Applied (if applicable):	

# Detecting Exoplanets from light curves of Kepler Mission using Fast Fourier Transform and Recurrence Plots

Aanchal Singh  
X19221771

## **Abstract**

The quest for finding life and habitat in the universe other than earth has been going on since long. To uncover it National Aeronautics and Space Administration (NASA) has Kepler Mission which captured data of brightness of stars which are called light curves. These are formed when a star transits from one place to another its host star. This star could be an exoplanet and this is decided by checking its brightness. This data is of a period of time and is in time series format. Using this data, Support Vector Machine is applied first. To improve the performance of models this research uses pre- processing techniques like Fast Fourier Transform and Recurrence plots are applied before feeding the data to any models. Then models like Support Vector Machine and Convolutional Neural Network with VGG16 are used. Different patterns were plotted and time series data was analysed to check if the data point is exoplanet or not. The results for with and without pre processing techniques are compared and they show how FFT and RP make a difference in getting improved performance in terms of accuracy, precision, F1 score, etc.

Keywords - Exoplanets, light curves, Astropy, Machine learning, deep learning, pre processing, Fast fourier Transform, Recurrence plots

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Related Work</b>	<b>4</b>
2.1	Researches using Transiting Method . . . . .	4
2.2	Methods to Extract Light Curves Data from Astronomical Data . . . . .	5
2.3	Pre-processing of Light Curves after Extraction . . . . .	6
2.4	Classification techniques using Deep learning . . . . .	8
<b>3</b>	<b>Methodology for Detecting New Exoplanets</b>	<b>9</b>
3.1	Data Selection and Extraction . . . . .	9
3.2	Transformation of Data . . . . .	10
3.3	Pre-processing Techniques . . . . .	11
3.4	Data Modelling . . . . .	11
3.5	Evaluation and Results . . . . .	11
<b>4</b>	<b>Design Specification</b>	<b>11</b>
<b>5</b>	<b>Implementation</b>	<b>12</b>
5.1	Data Cleansing and Exploratory Data Analysis . . . . .	13
5.2	Fast Fourier Transform . . . . .	14
5.3	Recurrence Plots . . . . .	15
<b>6</b>	<b>Evaluation</b>	<b>16</b>
6.1	Experiment 1: Replication of State of Art . . . . .	17
6.2	Experiment 2: Application of Recurrence plots . . . . .	18
6.3	Experiment 3: Fast Fourier Transform . . . . .	19
6.4	Discussion . . . . .	20
<b>7</b>	<b>Conclusion and Future Work</b>	<b>20</b>

# 1 Introduction

National Aeronautics and Space Administration (NASA) has many missions which assist in finding habitat other than earth. Among other different missions carried by NASA they have a critical mission of finding life beyond Earth. The Kepler mission uses Kepler Space Telescope which is used to discover more planets which are called exoplanets. Exoplanets are being discovered since 1960s. The Kepler project was designed to hunt for habitat and study planets that are similar to our own. There have been 2344 exoplanet discoveries which were detected through the transit method. Majority of exoplanets were detected by making use of indirect approaches like photometry and radial velocity. This research focuses on detecting exoplanets using light curve data. By making use of different pre-processing and machine learning techniques the transition of light curves is extracted into patterns and these can be used to discover an exoplanet around an alien star.

The data for the research is in Flexible Image Transport System (FITS) format. For plotting and processing of FITS file, which is time series data, this work uses the Astropy.io package in Python. Light curves are generated when a planet transits from one place to another. As they transit there are variations in light which creates light curves. These curves act as a proof towards being an exoplanet. A combination of features and FFT is used to acquire information from time series data. FFT has been widely used to analyze time series data and uncover patterns. This research entails observing several stars throughout time and assessing the results. To examine the light curves, Kalaei and Hasanzadeh (2019) used the Fast Fourier transform. Improvements in machine learning and numerical methods has made it possible to use them to learn and predict complicated patterns in a fixed amount of time (Woodward et al.; 2019a). Neural network provides 81 percent accuracy. Deep neural networks are also used for classifying light curves. The data is flattened into single dimensional vector and then it is fed to Artificial Neural Network (ANN) by Asif Amin et al. (2018). These two papers provide us the result that deep learning techniques provide better results.

Recurrence plot (RP) is an advanced technique to analyse non linear data. It is a visualization of matrices of distance between points of time series in a pseudophase space. RP unfolds the times when pseudophase of dynamic system visits same area in same space. Recurrence plots were used to convert the time series data into images to analyze and then apply the neural networks. Along with these the novelty in this research is using pre-processing techniques which can provide better results. Neural and fuzzy network was used to classify results with accuracy and find potential exoplanets by Basak et al. (2020). Chaushev et al. (2019) made use of Convolutional Neural Network (CNN) which was used for Next Generation Transit Survey (NGTS). Fully simulated data along with injected planetary transits and real data was used in order to train the data set. Shallue and Vanderburg (2018) had used a combined linear architecture of CNN which was a fully connected architecture to distinguish and classify the exoplanets. These researches used different neural networks but did not have proper pre-processing techniques to exoplanet data. In order to get better results pre-processing plays an important role. This research thus considers FFT as an important part of the solution of the below research question.

This research work focused on below research question:

*“To what extent using pre-processing techniques like Fast Fourier Transform and Recurrence plot along with machine learning and neural networks, improve prediction of exoplanets from light curves?”*

When there is a planetary transit happening between the star and earth a light curve

pattern is formed. The research objectives are as below:

- 1) Review different state of art literature to understand the domain better.
- 2) Pre-processing of light curves using Fast Fourier Transformation and Recurrence plot.
- 3) Apply neural network like Convolutional Neural Networks and machine learning model like Support Vector Machine.
- 4) Evaluation of the models and compare which pre-processing techniques helps in getting better results
- 5) Compare which model provides more accurate and precise results.
- 6) Using different evaluation metrics like accuracy and confusion matrix decide which approach classifies exoplanets in a better way.

The hypothesis for this research can be said that the light curve pattern under analysis is an exoplanet and the alternative hypothesis will be that the light curve is not an exoplanet. The novelty in this approach is using FFT and recurrence plot for pre-processing before developing the final model.

In this document, after this introduction, the document covers different literature review of related work in Section 2. The research methodology which is to be implemented from different related work is described in detail in Section 3 of the paper. The architecture design which is formed for implementation is covered in Section 4. Section 5 of the paper has different tools and technologies needed for the implementation of the model. The final outcome and conclusion of various experiments carried out is described in detail in Section 6. The aim of the paper along with final conclusion and future scope is discussed in Section 7 of the paper.

## 2 Related Work

Exoplanets are as interesting as their name and to understand them better and the approach to discover them, it is necessary to review some work in this domain. Getting the light curves from the data which is sent by telescopes is only the first step. We then try finding ways to work on these light curves. They are extracted and later processed to remove nulls or noisy data. To get accurate results researches where pre-processing techniques are used were cited. The classification of light curves has been done using different machine and deep learning techniques. To understand these models along with FFT and recurrence plots an in depth study of previous researches is needed. The study is divided into four parts as below:

1. Researches which have used parametric modelling
2. Methods to extract light curves data from astronomical data
3. Pre-processing of light curves after extraction
4. Classification techniques using Deep learning.

### 2.1 Researches using Transiting Method

The transit method is the most prevalent method initially to detect exoplanets. It checks for the planets moving from a star and an observer. They form light curves which help in determining if it is an exoplanet or not. Transition of light curves are time series data which are being observed from many years and as they move from front of a star the lights dim and also there can be noise in the data. Carado and Knuth (2020) created an algorithm which helped in checking the transit light curves and model them for exoplanets. They

checked the cross sectional area of the planet and also difficulties which can be seen in arbitrarily checking a random orientation in space. The co ordinates of a random point on light curve were taken and the ellipsoid was modelled.

Ferrari et al. (1999) used the COROT data and parametric modeling to detect an exoplanet. They record photometric signal and then calculate ratio of planet to the star surface. The transit creates a jump in mean and variance and this is captured by using Neyman-Pearson detector. If the value is lower than critical value then null hypothesis will not be considered. When null hypothesis cannot be considered then there is rise in statistics. This rise was caused by the alien planet.

de Almeida et al. (2020) looked for transit signals which were similar to that of Earth. They applied DT algorithm to the Kepler's TESS (Transiting Exoplanet Survey Satellite), CoRoT and Planetary Transits and Oscillations of stars (PLATO) light curves. They could eradicate the noise and create visuals which helped in better analysis of the data.

Light curves of bright stars was used by Southworth et al. (2020). To remove the noisy elements from the data they ignored the third light effect. The third light is the light which constitutes the part of entire light but it comes from a third body. This light will be of no use in our data and hence they identified this and removed it. After this they plotted the clean light where they made use of fractional radii of the planet and the host star. Even after this there could be some contaminated data which could be ignored.

## 2.2 Methods to Extract Light Curves Data from Astronomical Data

Different planets may have their own atmosphere and could have traces of water and hydrosphere. Santos et al. (2019) tried finding water on these exoplanets by modelling hydrosphere of Earth. They then tried finding similar patterns with the exoplanets. To check this they used ASTROPY and SCIPY. By making use of Lyman spectroscopy they used these libraries to check the velocity of neutral hydrogen which could be present in upper atmosphere.

Spectroscopy and Doppler imaging was used by Sluijs et al. (2019) for detecting exoplanets from transiting exoplanets. Using this technique they were able to remove the usual process of star observation. When the exoplanets transit from host star there are distortions that happen. Due to this there is a need of finding reference star measurements. All these steps were carried out by making use of ASTROPY. By making use of Doppler effect the noise present in the data was reduced. This method did not use any calibration observations for finding exoplanets.

The data set from Mikulski Archive for Space Telescope (MAST) was used by Woodward et al. (2019b) for looking at the observations of different Kepler objects. The FITS file present in this data set took 12393 Kepler objects under observation. The classification of these light curves was done by selecting the different headers and data units. To extract the light curves for classification ASTROPY was used in this research. ASTROPY consists of a sub package called PyFITS project - astropy.io.fits. This sub package was used by Barrett and Bridgeman in the year 1999. By utilising the different command line scripts they extracted the header and data from the FITS file. The commands used included `getheader()` and `getdata()`. The light curves were extracted from FITS file and it was made ready for pre processing of the data.

## 2.3 Pre-processing of Light Curves after Extraction

In order to apply any machine learning or deep learning approach to a data set it should be pre processed to provide better results. Pre-processing of data plays an important role building a model. In this research the FITS file which is in time series format is pre processed using Fast Fourier Transformation(FFT). The data from FITS file is in time series format. Fast fourier transform will decompose so that we get frequency from the time domain. The main aim for FFT is looking for any periodic functions. It is a sum of cosine and sine functions from different features involved. Another way of pre processing the data which is used is the Recurrence plots which help in classifying images of the light curves. The time series data is transformed to two dimensional images to check the recurrent characteristics of the time series data.

The need of pre-processing data was explained by Bugueño et al. (2021). The parametric models need to take different light curves. Using different pre processing techniques they tried to showcase how pre processing helps in improving the prediction results of different algorithms. In their research they had used different machine learning techniques like logistic regression, random forest and Support vector machine (SVM). After applying pre-processing technique they evaluated the models on the basis of F1 score.

Recurrent Neural networks were used by Hinners et al. (2018) to classify light curves. The results that they received were not up to the mark. To improve their results they studied further and found that features for classification and regression problems can give better results when pre processing is done to the light curves. A pre processing technique called Synthetic Minority Over sampling techniques (SMOTE) can give much better results to classification problems with an improvement in the accuracy of how the data is classified. They used SMOTE and post that applied RNN again. A considerable difference was observed in the results when the accuracy was being considered. By this research it can be concluded that light curves give much better results when pre processing of data is done. They also compared different models and their performance after applying SMOTE. They checked the Eout value which stands for standard error and found thier expected and observed values. These were found for different machine learning algorithms and were applied without any pre processing which means the data was raw and noisy. Below table 1 shows different accuracy and error values which were observed in this research.

Table 1: Errors and Accuracy values for different algorithms

<b>Classifier</b>	<b>Eout</b>	<b>Accuracy</b>
Decision Tree	8.19	62.82
Random Forest	4.64	57.58
KNN	4.72	58.96
SVM	5.71	57.72
Naive Bayes	92.38	51.15

To understand the behavior of light curves FFT was used by Kalae and Hasanzadeh (2019). A study was carried over the Scuti star from a period of 1970 to 2017. The time series data helped in observing the behavior of the star. They had segregated the light curves to alternative deep and shallow lights which have irregularities. To determine important features FFT and PSD was used. The behavior of this star difficult to reckon and time series data helped in comprehending linear stochastic process with a resonance



between the fundamental mode and the first overtone. FFT showcased non linearity effects which were there and helped in investigating the behavior. This was validated later by using power spectral decomposition. FFT helped them uncover various power spectrum of signals created by RSct over the span of all the years under consideration. The two techniques helped them to determine that there are two dominant peaks in the curves. These curves imply alternating amplitude and frequencies which show that the curves are not constant. This variation was observed in the time span of 10years. The second peak was observed as a quasi-periodic behavior.

FFT was used by de Francis et al. (2019) in order to understand the data over a period of time. Light curves of Scuti stars were taken into consideration which were observed for a long period of time by the CoRoT satellite. Fraction of these light curves were taken for a fractal analysis. Using this analysis an attempt was made to understand the fluctuations in the curves. FFT helped them understand the fluctuation of stars which was assisted by pre whitening technique. This helped in stabilizing the data of Scuti stars. FFT played a very important role in this research for studying the time series data.

The importance of pre processing techniques was also verified in the research by Huo (2018). To study the time series data he used two pre processing techniques. One being FFT while the second be Data Time Warping(DTW). The time series data used by him was used to forecast electric loads. The machine learning techniques which he used were Support Vector Machine and Random Forest. If there is noisy data or erroneous data these models do not provide accurate results. Also the features cannot be identified. Hence a combination of DTW and FFT was used for processing the raw data. The data was taken from NYISO which was for a period of 15 years. DTW helped in finding similarity while FFT helped in finding reference parameters and focused on periodicity and relative strengths of any components. Normalizing the data and smoothing of the curve was faster after applying these techniques.

Sartoli et al. (2020) classified malware binaries as per their occurrence and recurrence. A stochastic process was carried out and recurrence plots were used to convert the binary values of malware emissions into two dimensional texture images. Post which CNN was applied which classified malware images into families. A comparison was made of the performance of CNN by using recurrence plots and when these images were directly transformed from malware binaries. The CNN used three convolutional layers of two dimensional filters to smooth the activation maps.

Time series data was classified in two ways by Kirichenko et al. (2020). They considered quantitative fractal characteristics as classification and compared them with recurrence plots. Recurrence analysis has its basis on how often the data is repeated and how it is presenting the data as a geometric structure. The diagonal structures show the time interval when time series section are close to each other. Post this a simple feed forward neural network with ReLU activation function was used.

The analysis of High Resolution melting data where the melting temperature and melting curve classification was done by Ozkok and Celik (2021). The HRM data is vital when detecting diseases like Covid-19, for checking spoiled food. In order to improve the accuracy they put forward the idea to visualizing the data using recurrence plots and then apply CNN to these images. Grey scale recurrence plots were created and performance of this model was evaluated which gave improved accuracy of classification.

## 2.4 Classification techniques using Deep learning

After the pre processing of data is done, the data is ready for application of machine learning or deep learning models which will help us in classifying the exoplanets. This will help us in getting if the light curve is exoplanet or not. We can evaluate the results we get after this step. This part of related work will help us understand how light curves are predicted using different machine and deep learning algorithms.

In the initial days, exoplanets were detected by using transit method. These curves are formed when planets move away from its host stars and form these curves. Chintarungruangchai and Jiang (2019) used convolutional neural network for prediction of light curves. A combination of Artificial Neural Network along with convolutional neural network was used by them. In addition to these they also used logistic regression. Spatial data is taken into consideration where time series played a major role. Stochastic gradient descent helped in creating training data and which assisted in finding the major features in this process. CNN provided better accuracy and precision as compared to traditional transition method.

Linares et al. (2020) used CNN to detect new planets from light curves. The shape and mass of different exoplanets was taken into consideration here to determine these space objects on the basis of their physical features. Machine learning algorithms were compared with CNN. Out of the two algorithms Bagged trees was an ensemble technique while the second was SVM. To model the light curves bidirectional reflectance distribution function (BRDF) was used. BRDF separated the data into two parts: specular component and diffuse component. Real light curves were used which were taken from Multichannel Monitoring Telescope (MMT). After this different space objects like debris, rocket and satellite bodies were put into three different categories. Learn-able filters were applied to filtered images from core layer to find features. Weights are added as desired so that different images come to a single point. The images were found and were divided, then multiple values were added to original pixel values. ReLU function was then used to add filters. The training of the model was done for 8000 scenarios which had 9 space objects and which had 2000 epochs. The accuracy of this model was achieved as 97.83

The periodic transit of light curves was studied by Bugueño et al. (2018). The periodic transits were checked if they were candidate planets and they were not false positives. Radial velocity was a way of looking these planets. They took the mean, deviation and kurtosis for cleaning and pre processing of the data. Principal Component Analysis was done to find main features of the data and later SVM and logistic regression was applied. After this fully connected CNN was applied with astronnet process where final outputs gave a probabilistic prediction of the data.

Jagtap et al. (2021) created an architecture where they look for life on exoplanet by using features like gravity, eccentricity, mass, radius and other characteristics. They considered different machine learning algorithms to classify exoplanets in different classes of habitat. They used KNN, Random forest and SVM and then evaluated these models on the basis of accuracy and precision. They applied SMOTE as a pre processing technique followed by normalizing, smoothing and later standardizing the data. Once these are done they train the model on ASTRONET architecture which is similar to CNN. They use the sigmoidal activation function with max pooling which reduces over fitting of data.

Bu et al. (2019) used CNN and SVM together to find hot subdwarf stars. LAMOST telescope was used to get the data for this star. Data for five years was collected and classification was done after eliminating the noisy data. As the features are not linear

SVM solves the issue by using kernel function.

Throughout all the papers mentioned it was discussed how transiting methods were used to classify exoplanets before machine learning techniques. Different researches gave us an idea how pre processing techniques have improved the prediction models with different stars and data with different satellites. FFT and Recurrence plots are proved to improve results with electric load and malware and melting data respectively. This research focuses on using these two techniques on light curve data. Previous researches

All papers mentioned above discuss how pre processing techniques have helped in improving the model accuracy and precision. Models like CNN and SVM are used to classify light curves and astronomical data. This section various papers where relevant researches are discussed and are critically reviewed.

To summarise the critical review of study, transit method used different techniques like parametric modeling and different algorithms to detect algorithms. In order to extract light curves data of satellites PYTRANSIT and ASTROPY are used by researchers along with PyFits. The pre processign techniques used used by Hinners et al. (2018) was SMOTE upon which different alogrithms were applied with showed considerable improvement in classifying the data. Scuti star data from 1970 was applied with FFT which showed the behavior of stars over the years. Recurrence plots were not used on light curves by Ozkok and Celik (2021) but on Covid-19 data and get results using CNN. Classification of light curves and the images of them was done using various machine and deep learning algorithms. Activation function like ReLU was applied with different epochs. After applying SMOTE machine learning algorithms like KNN, Random forest. SVM were used.Keeping these researches in mind and the novelty factor to be implemented, recurrence plots is applied to be followed by CNN and FFT is applied followed by SVM in this research.

### 3 Methodology for Detecting New Exoplanets

To classify FITS data into exoplanets there are a number of stages the data has to go through. To keep a note of these different stages, a standard procedure has to be followed. The methodology used in this research was Knowledge Discovery in Database(KDD). KDD proves to be one of the best methodologies to be used with neural networks Bing (2017).

In the section below a detailed description of how this process was implemented is given. Classification of these stars will be as: 1) Planet Candidate: A confirmed planet which transits from host star. 2) False positive: Any space object which travels from planet and host star. 3) Phenomenon of non transiting: This may happen due to error with the equipment. fig:figureExample1 shows the complete process of how exoplanets planets are detected. Figure 1 helps in understanding the methodology followed in detecting exoplanets.

#### 3.1 Data Selection and Extraction

The time series data was taken from Mikulski Archive for Space Telescopes (MAST) <sup>1</sup>. The data was in time series format and needs to be analysed to understand if it is a planet or not. Data can be downloaded in order to get patterns from over the time and

---

<sup>1</sup>Kepler Mission Data: [http://archive.stsci.edu/kepler/data\\_search/search.php](http://archive.stsci.edu/kepler/data_search/search.php)

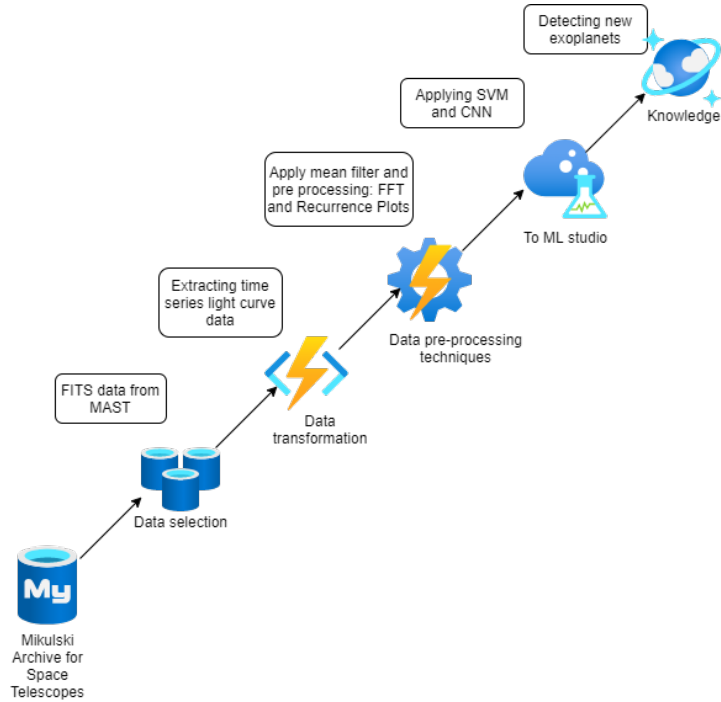


Figure 1: Exoplanet Detection using KDD

train the models to get accurate classification results. Using this portal, links to light curves data of all the years was obtained as a text file. Links for both exoplanets and non-exoplanets were taken and stored in variables. This data being in FITS format was extracted using ASTROPY. The data once extracted was saved in CSV format for better analysis.

### 3.2 Transformation of Data

The file contains header information and data related to luminosity and flux. Every light curve can be accessed by taking into consideration 'TIME', 'SAP FLUX' and 'PDCSAP FLUX'. These features are stored in variables. It was then converted to a tabular format. This data had null values which were eliminated by using statistical mean of the data. The flux data was then converted into a tabular format and exploratory data analysis was done to understand the features in a better way. Using the matplotlib.pyplot library of Python this step was completed. Along with this Pandas was used to convert the FITS file to readable format and also removing null values and replacing them with their mean values was done. The data being time series has long set of patterns. To understand the data and have an optimum model, a part of time series was chosen. Pre processing was done on this part. To have an aligned set of rows each row of the selected data had its median value set, for the desired length of the data. Image resizing and reshaping was also a part of this step. The images for Recurrence plot were gray scale and were huge images. These were scaled down to feed them to the model.

### 3.3 Pre-processing Techniques

Once the data was in readable format in order to improve results by our model, the novel approach of applying FFT and Recurrence plots was the next step. These two techniques are applied by using Scipy and matplotlib libraries. For FFT the scipy has fftpack library which has fft and fftshift. Recurrence plots are built on a subset of main data set. The plots are in 1-D format at thwas stage. Classification of images was done as per positive and negative cases. It includes scaling and converting the images into 2-D and 3-D images so that they can be fed to the models. The data considered for recurrence plots was 34744 links for exoplanets and 50002 images for non exoplanets. A part of this data of 8000 exoplanets and non exoplanets was considered. 20000 data points of light curves are considered for pre processing.

### 3.4 Data Modelling

This step involves model training on the split dataset of 6400 and test set of 1600 data points. The pre processed data for FFT was trained with SVM while data with recurrence plots was trained by using CNN. To get better results from recurrence plots pre trained model VGG16 was trained with image shape (100,100,3). Curves are flattened and model was trained with 50 epochs and ReLU and softmax activation function. To optimize the model Adagrad optimizer was used with categorical crossentropy loss function. The post training function was used to get small sized model and to also get a good performing model.

### 3.5 Evaluation and Results

The performance of both machine learning and deep learning models before and after applying pre processing techniques was evaluated. The time series data on application of FFT was evaluated on the basis of accuracy, F1 score, precision and recall. Same data was converted to images and on the basis of same parameters the model was evaluated. The optimal model was chosen after different experiments and the final result was interpreted.

## 4 Design Specification

The design section architecture and flow which is followed to discover exoplanets. We will first see how the model was built in Figure ?? which follows a bottom up approach for classification of exoplanets. The data layer shows how FITS file was processed using different components like ASTROPY, FFT, Recurrence Plots, SVM and CNN. The Data layer takes FITS file from MAST portal and then they are processed using Astropy, Pandas, Keras, Scipy and Tensorflow libraries. Cleansing of data by using statistical functions like mean and median was done. Once the null and noisy data was cleaned pre processing was done using FFT and Recurrence plots. The second layer is application layer where the pre processed data was applied to two models. SVM for FFT processed data and CNN for recurrence plots processed data. A pre trained model of VGG 16 with different dense layers was applied. The last layer was where the evaluation was done and is the Presentation Layer. Different metrics like accuracy, precision, recall, F1 Score and ROC curve are taken into consideration to evaluate and compare the models and visualize the results. This is the Presentation layer of the architecture.

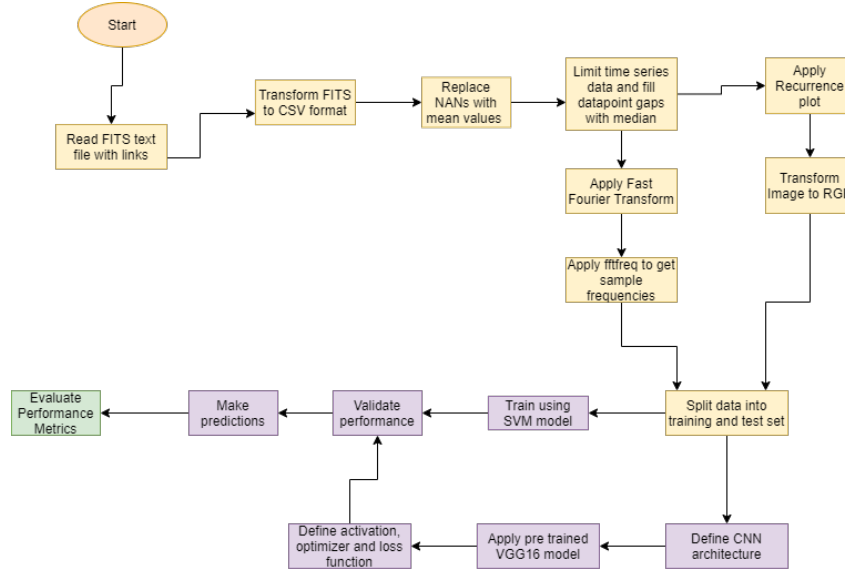


Figure 2: Architecture for Classifying Exoplanets

It is essential to know how classification of exoplanets is done from the raw data. The flowchart in Figure 3 explains step by step how FITS file was getting segregated into exoplanets and non exoplanets.

First values are taken as exoplanets and non exoplanets from NASA’s MAST portal. Converting this into CSV format, cleansing of data was done. FFT and Recurrence plots are applied on this raw data. Recurrence plots need to be converted to RGB format before they are applied with any model as CNN requires 2D or 3D images. They are stacked and resized so that CNN can be applied on them. The data was then split into training and test sets so that the model can be trained.

FFT processed data was directly applied with SVM while Recurrence plot data was applied with VGG 16 where default top layers are replaced with customized dense layers. Neural networks being stochastic gradient descent algorithm refer to error gradient which helps the model in predicting. To optimize the performance and minimize the objective function, loss function was used here. All different numbers are brought down to one scalar value. Error between probability distributions was measured by using cross entropy. While classifying the data as exoplanet or not the probability of light curve being an exoplanet may be predicted. Cross entropy will be when a set of model weights was looked after, where difference between the model’s predicted probability distribution and the distribution of probabilities in the training dataset was considered.

Once the models are applied, both are validate on their performance. Predictions are made into two classes as positive case: Exoplanet or negative case: non exoplanet. Evaluation of these predictions was done on the basis of metrics like accuracy, precision, F1 score, etc.

## 5 Implementation

The flowchart in Figure 3 is a step by step process of implementing the project. The first step of implementing this project will be extracting data from the FITS file obtained from MAST portal. For this different python libraries like Astropy, Keras, Tensorflow,

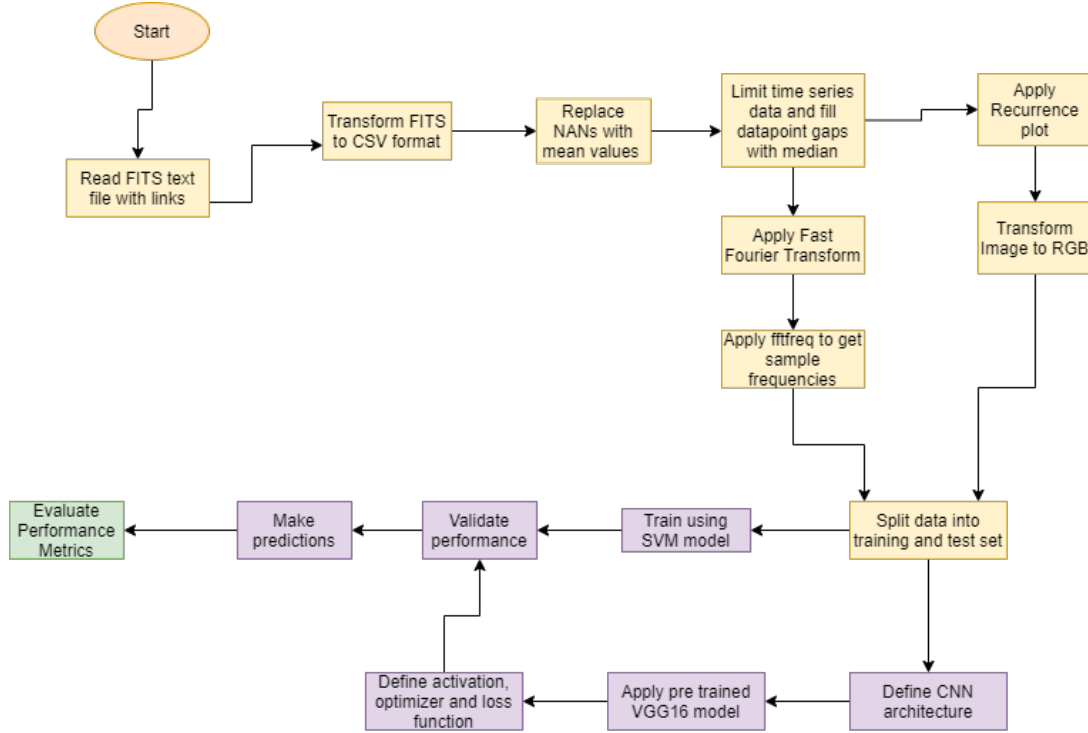


Figure 3: Flowchart for Detecting Exoplanets

Scipy, Sklearn, Numpy and matplotlib are used. Astropy helps in extracting FITS file while Scipy helps in applying FFT on the data. Tensorflow helps in implementing VGG16 model and CNN. All these files are imported in the project.

## 5.1 Data Cleansing and Exploratory Data Analysis

The FITS file are in the form of links are are downloaded in text format. To use this data further the test file was first cleaned. All NAN values are replaced by the mean of the data. PDCSAP.FLUX is the feature which interprets the brightness of the star. After taking the mean these are converted into CSV files and also converted to dataframe. The data being time series has huge number of data points. This research limits the data points to 8000. Each light curve was taken upto 20000 rows for these points. It is not necessary that every light curve will have these many rows and hence the data points having NAN values were first replaced with 0 and later were replaced by the median values of the data.

After initial clean up of data both exoplanet and non exoplanet data was divided into training and test sets before doing the Exploratory Data Analysis(EDA). The time series data was plotted using the flux values for both exoplanet and non exoplanet data. The plots for exoplanet and non exoplanet showcase patterns for both of them.

It can be observed that exoplanets have a pattern over the period of time. Non exoplanet data does not have any pattern and have random flux. The data was divided into training and testing sets. 80 percent of curves was trained while 20 percent was tested. SVM was used to train the model without any pre processing techniques. The results are discussed in detail in evaluation section.

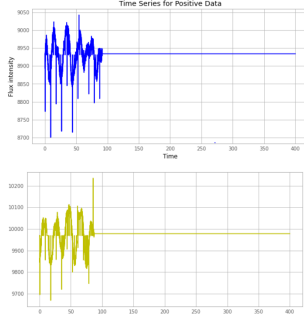


Figure 4: Exoplanet

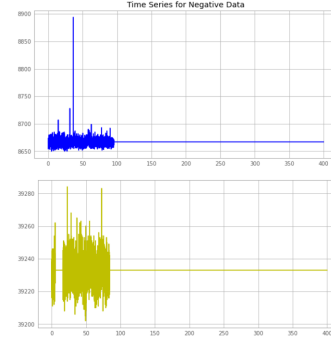


Figure 5: Non Exoplanet

## 5.2 Fast Fourier Transform

To get accurate and precise results from the model, it was important to use pre-processing techniques. FFT was applied on the cleansed data, on both training and testing data sets. For FFT the Scipy and numpy packages in Python are required. It will take a sum of all periodic components, in this case PDSCAP\_FLUX, to recover signal from these components. This signal was truncated and a rectangular window function was used for sampling the data. The sample size here was declared in a variable `f_s` variable as 100. `fftpack` is a part of `scipy` which has `fftfreq` function which returns the sample's frequencies. The final array of FFT data will have centers of the data with sample spacing. A window of length `n` and sample size `d` works in following way when FFT was applied.

$$f = [0, 1, \dots, n/2-1, -n/2, \dots, -1] / (d*n) \text{ if } n \text{ is even}$$

$$f = [0, 1, \dots, (n-1)/2, -(n-1)/2, \dots, -1] / (d*n) \text{ if } n \text{ is odd}$$

In this research the sample size was taken as the length of training data set and testing data set while `d` has the value of 100 as defined in sample size. Plotting these values using `matplotlib` the following graphs are seen for exoplanet and non exoplanet data. The X axis has the frequency of data while Y axis represents the light spectrum to calculate the magnitude.

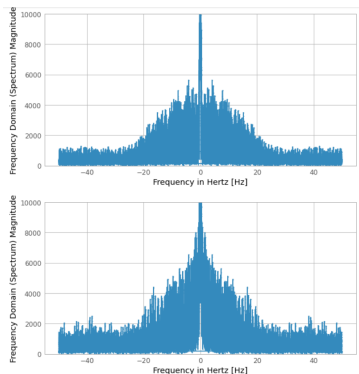


Figure 6: FFT Analysis for Exoplanet

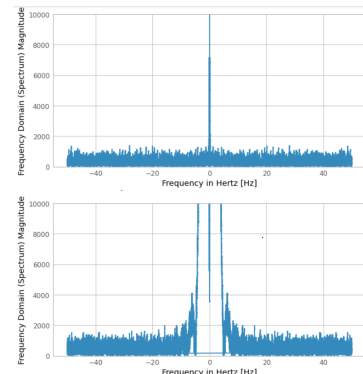


Figure 7: FFT Analysis for Non Exoplanet

The plots show that the data with exoplanets have a considerably high magnitude of spectrum for most of the light curves and they form a pattern. The data points of non exoplanets have short spectrum and only a few curves have high spectrum and they do not form any pattern whatsoever.



### 5.3 Recurrence Plots

Recurrence plot visualizes data of a matrix where recurring data was plotted and revealed on a pseudo space. The recurrence plot of exoplanets was plotted on a 1 Dimensional space where recurrence of a state at a time  $i$  at a different time  $j$  was marked within a 2D squared matrix with ones and zeros represented as black and white dots. Both these axes represent time. To plot the time series pdist function was used from scipy which gives distance between observations in the pseudo space. It takes the array of time series data where the training data set was given with no particular metric. It returns a condensed matrix where for each  $i$  and  $j$  the distance was calculated. The values are then rounded up for each sample. This output was in condensed matrix format which can be converted to square distance matrix by using squareform function from scipy library. It returns a redundant matrix which can be used plot the recurrence plots.

The recurrence plots built for the light curves in this research are as below. The plots

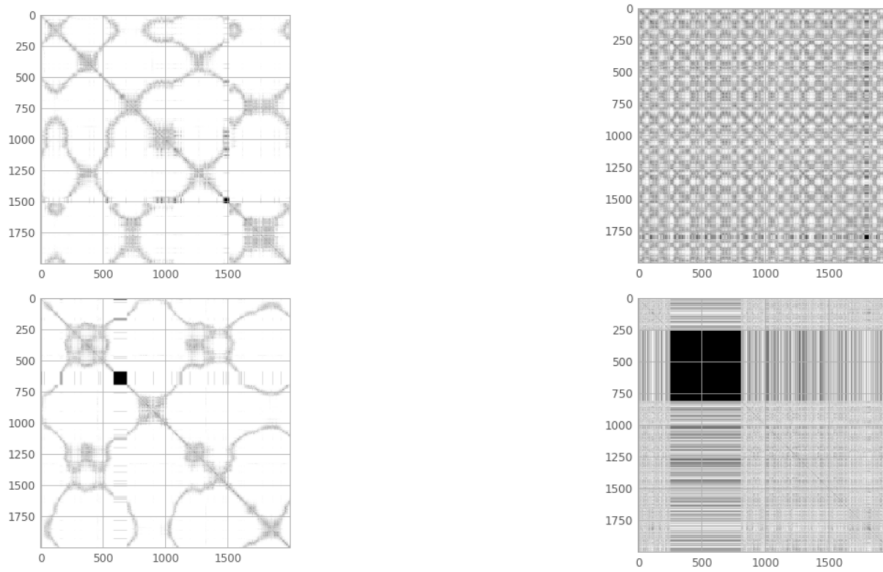


Figure 8: Recurrence Plot for Exoplanet    Figure 9: Recurrence Plot for Non Exoplanet

for exoplanets represent oscillating systems with diagonal oriented and periodic recurrent structures for the time duration. These patterns infer that they are exoplanets. The plots for non exoplanets have a drift, single and isolated RP. They show heavy fluctuation. It represents uncorrelated random process which signifies that they are not exoplanets.

To apply model on data pre processed by recurrence plot the images must be either 2 dimensional or 3 dimensional image. As the images are very huge they need to be reshaped and resized. They are to reduced the size to height and width of 100. Using CV2 package images are resized. Once the size of image was reduced it was necessary to change its dimension and that's when Numpy's stack function was used. This function helps in joining a sequence of same dimension along a new axis. It takes an array of sequences of the same shape with the axis of the resultant array along which the input was to be stacked. It will return stacked array having more than one dimensions. After stacking the images look as below. These images also infer that exoplanet image is systematic and has some pattern while the non exoplanet is random data spread throughout the pseudo space. These images were then ready to be applied with CNN. The model uses a pre trained model VGG16 where 16 weight layers are used (13 convolutional layers, 3 Fully Connected layers).

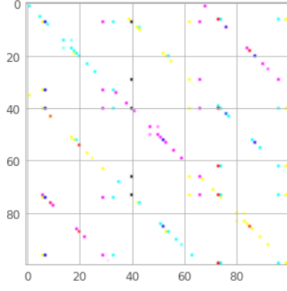


Figure 10: Stacked Plot for Exoplanet

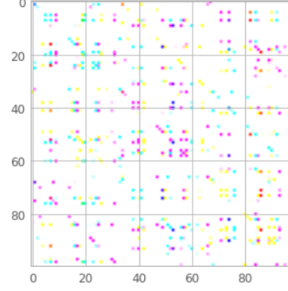


Figure 11: Stacked Plot for Non Exoplanet

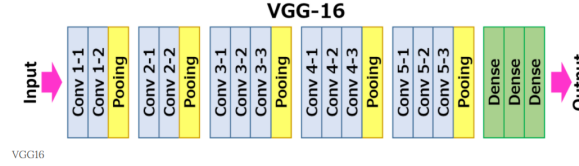


Figure 12: VGG 16 Architecture

The input was a 100 X 100 image followed by mentioned structure. The number of input nodes and hidden layers can be easily changed. In order to remove bias balanced weights are used by customizing them. Customized flattened layers are used with activation function with hidden layers as 256, 180, 128 and 64 ReLU activation function. Loss function was introduced to every prediction target where actual probability of the class was present with predicted probability. Customized 4 Dense layers are applied on base layer of the model followed by prediction layer which uses Softmax activation.

In normal gradient descent the learning rate was fixed. The model was taking too long for training when it was applied to the images of light curves. Hence optimizer was required to bring the learning rate of the model to an optimal point. Before using Adagrad optimizer RMSprop and Adam optimizers were used. Adam stores the learning rates along with the momentum. It was giving less accuracy for this project. Hence RMSprop was used which decreases the learning rate and which was comparatively slower. When learning rate was more and optimizer tries to get minimum learning rate. It either oscillates to reach minimal learning rate or takes small intervals to reach the optimal value. Using weights this can be done and RMSprop uses sliding window which decreases the sum. Hence in the end to have an optimal result with which was also quicker Adagrad was used. The accuracy of 54 percent was same as RMSprop but Adagrad was quicker than it and hence Adagrad was used for detecting exoplanets and non exoplanets.

## 6 Evaluation

This research considers the data and methodology as baseline which was used by Hinners et al. (2018). They pre processed 94GB data from MAST. Cleaning of data for NAN values and checking the plots for patterns and deciding if they are exoplanets are not was a part of their research. They also used VGG 16 with optimizers, loss function and cross entropy. The authors have their stand that LSTM RNN classify the light curves with less accuracy. There is noisy and sparse data with different models like KNN, D-Tree and Random forest. The main motive of this research was using pre processing techniques

```
model.summary()
```

Model: "sequential\_1"

Layer (type)	Output Shape	Param #
vgg16 (Functional)	(None, 3, 3, 512)	14714688
flatten_1 (Flatten)	(None, 4608)	0
dense_7 (Dense)	(None, 128)	589952
dense_8 (Dense)	(None, 64)	8256
dense_9 (Dense)	(None, 2)	130

Total params: 15,313,026  
Trainable params: 598,338  
Non-trainable params: 14,714,688

Figure 13: Model Summary

and improving the classification prediction and performance for different models.

To evaluate the model following metrics are considered in this research:

1) **Accuracy**:  $(TP + TN)/(P + N)$  where TP- True Positives TN- True Negatives  
P- Total number of positives N- Total number of negatives

A classifier with binary classification will have an accuracy of 1/2. The data for this research is balanced and has almost equal number of positive and negative cases. The accuracy will therefore will not be misleading.

2) **Recall**: The recall value of a model explains the number of positive cases which were identified correctly by the model.

Recall:  $TP/(TP + FN)$

3) **Precision**: The value of precision is the number of predicted positive cases which were predicted correctly by the model. It is represented as-

Precision:  $TP/(TP+FP)$

4) **F1 Score**:After having values for Recall and Precision, F1 score can measure the accuracy of the model. It considers both values which we get from the truth table. It is harmonic average of precision and recall values. The result is considered best when it is 1 and worst when it is 0.

F1 Score:  $2 * (Precision * Recall) / Precision + Recall$

5) **ROC Curve**: This is a graph which plots the true positive rate and false rate and shows the performance of the model at all thresholds.

TPR:  $TP/(TP + FN)$  FPR:  $FP/(FP + TN)$

This two values are plotted on the axes. Classification of items as positive is more when classification threshold is lower. As a result both false positives and true positives increase. The metrics in this research are validating how pre processing techniques can improve the performance of any model.

## 6.1 Experiment 1: Replication of State of Art

In Experiment 1, the experiment by Hinners et al. (2018) is replicated. Light curves are taken as a baseline on the basis on their flux value which indicates the brightness of the star. The time series data is taken and plotted on the basis of PDCSAP FLUX. Instead of directly feeding time series to VGG16 convolutional architecture after cleansing, like in the state of art, the data SVM is used. The results of classification give an accuracy of 52.15 percent. Evaluation Metrics for SVM without pre processing technique is as below:

The 0 here represents non exoplanets and 1 represents exoplanets. It is observed that accuracy is 52 percent with recall for positive cases being 70 percent and negative cases

```

accuracy_score : 0.52125
classification report :
      precision    recall  f1-score   support
0         0.53      0.34      0.42       86
1         0.52      0.70      0.59       86

 accuracy
macro avg      0.52      0.52      0.50      166
weighted avg   0.52      0.52      0.50      166

```

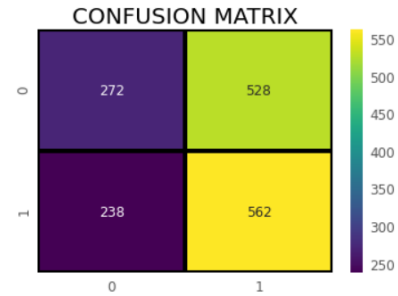


Figure 14: Matrix for SVM

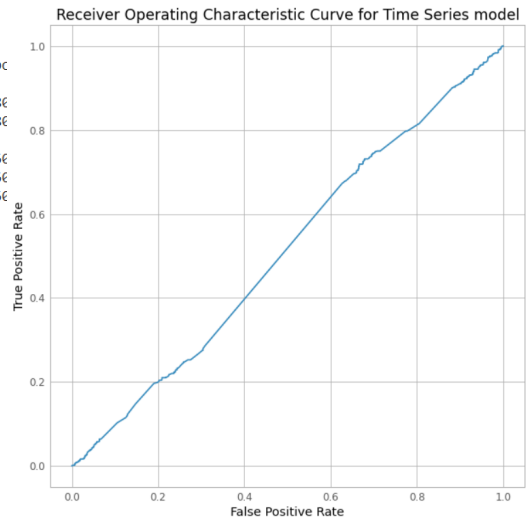


Figure 15: ROC Curve for for SVM

being 34 percent. Precision is 52 percent for positive cases and 53 percent for negative cases. the F1 score is 59 percent for exoplanets while for non exoplanets it is 42 percent. It can be observed that exoplanets are being predicted appropriately on a higher scale. ROC curve here signifies that 50 percent both cases are being predicted correctly.

## 6.2 Experiment 2: Application of Recurrence plots

Recurrence plots use images for classifying the data as exoplanet or not exoplanet. CNN with VGG16 improves the prediction. The evaluation metrics after using RP are as follows. It can be observed that there is slight increase in accuracy of the model after applying recurrence plots to the light curves. It is 54 percent now with a precision value of 54 percent for both cases. The recall value and F1 score is slightly less for exoplanet cases. The ROC curve also signifies that the exoplanets and non exoplanets are being predicted correctly for half of the times with the curve bending more towards False positive rate.

```

accuracy_score : 0.539375
classification report :
      precision    recall  f1-score   support
0         0.54      0.57      0.55       800
1         0.54      0.51      0.53       800

 accuracy
macro avg      0.54      0.54      0.54      1600
weighted avg   0.54      0.54      0.54      1600

```

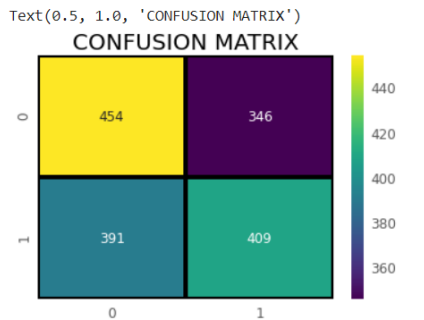


Figure 16: Matrix for RP

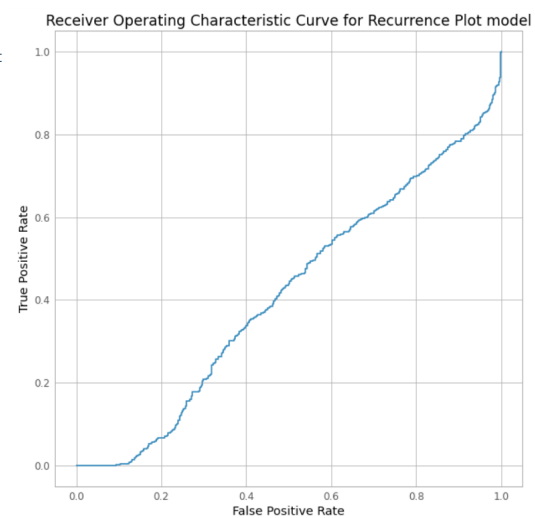


Figure 17: ROC Curve for for RP

### 6.3 Experiment 3: Fast Fourier Transform

Fast Fourier Transform being a pre processing technique for time series data increases the evaluation metrics considerably. The accuracy of the model SVM increases to 59 percent after applying FFT. It is 89 percent precise for identifying correct exoplanets. Confusion matrix signifies that 782 exoplanets are guess correctly while non exoplanets are predicted correctly for 642 cases. There is a slight bend towards TPR in the ROC curve.

Figure 20 shows the comparison of results of SVM before and after using FFT. It clearly shows that the true positive rate is better after applying FFT to the model. Previously there was no significant difference as the model was predicting almost equally for both the cases.

accuracy\_score : 0.586875

classification report :

	precision	recall	f1-score	supp
0	0.55	0.98	0.70	86
1	0.89	0.20	0.32	86
accuracy			0.59	166
macro avg	0.72	0.59	0.51	166
weighted avg	0.72	0.59	0.51	166

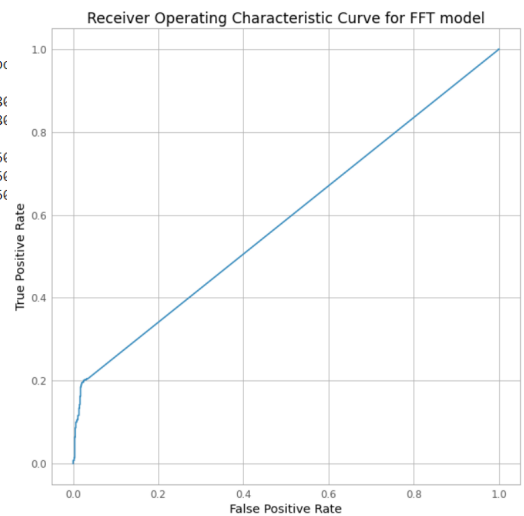
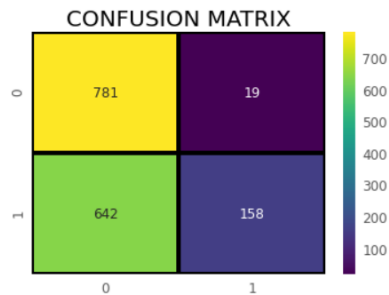


Figure 18: Matrix for FFT

Figure 19: ROC Curve for for FFT

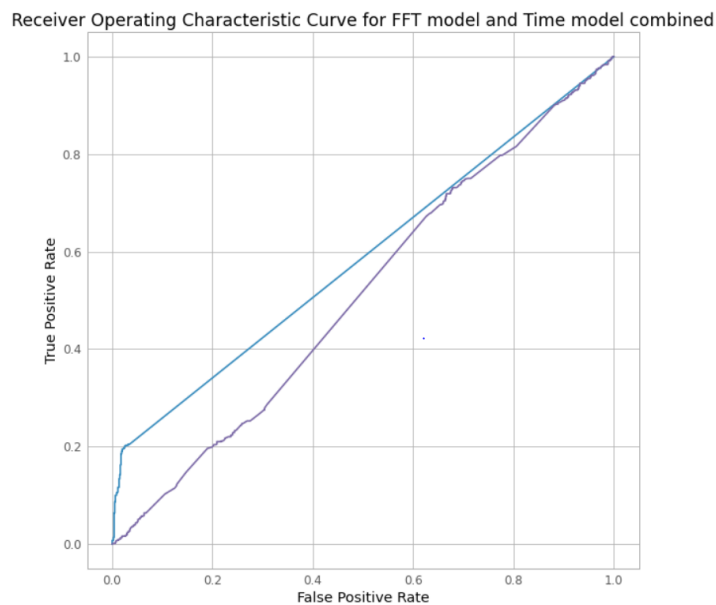


Figure 20: ROC for SVM before and after FFT

## 6.4 Discussion

All three experiments signify that the models perform better when applied after pre processing techniques. The classification is improved to an extent after FFT and RP. Two different algorithms were used here for assessing. SVM before application of FFT was predicting exoplanets with an accuracy of 52 percent which increased to 59 percent. Similarly, CNN is applied with Recurrence plots increases the accuracy of classifying data to 54 percent.

The purpose of this research of analysing the light curves time series data with pre processing techniques was fulfilled. The models selected for testing this approach gave satisfactory results. The accuracy of the models is not high as it would require huge amount of astronomical data for achieving a good accuracy. Considering the small sample that is taken in this research the model was trained and processed effectively and efficiently. Less data was used to save storage, memory cache and less time in training the model. Seeing these different factors the novelty of this research proved to be correct with increasing the performance of the model after applying pre processing techniques.

## 7 Conclusion and Future Work

Going back to the research question of whether pre processing techniques improve the prediction or not, was analysed in depth and reached a conclusion that there is improvement in results by both techniques. With all objectives like applying preprocessing techniques: FFT and RP, applying different models like SVM and CNN, evaluating their performances and in the end comparing how each technique and model perform are fulfilled. The findings of the research clearly re instate that pre processing techniques do have an effect on the models and improve their performance, Without FFT the accuracy of SVM was not finding exoplanet data correctly. After FFT was applied the positive data was predicted more precisely, Similarly after apply RP, CNN with VGG16 gave better predictions for exoplanets. The experiments performed clearly show the importance of FFT and RP. Without these techniques the models were not able to predict the patterns correctly for the time series. The limitation of this research was that it considered a small sample of the huge data, Using a larger sample could help in getting more accurate results. It also had an over fitting issue.

Astronomical field has many new researches ongoing. In order to getting accurate and precise results with exoplanets it is required that huge datasets get covered in the study. Time series data needs more analysis. Future researches can consider using more data for more accuracy and less time processing. They can also consider using Recurrent Neural Networks with Long Short Term Memory (RNN - LSTM) for better results and analysis.

## References

- Asif Amin, R. M., Talha Khan, A., Raisa, Z. T., Chisty, N., SamihaKhan, S., Khaja, M. S. and Rahman, R. M. (2018). Detection of exoplanet systems in kepler light curves using adaptive neuro-fuzzy system, *2018 International Conference on Intelligent Systems (IS)*, pp. 66–72.
- Basak, S., Saha, S., Mathur, A., Bora, K., Makhija, S., Safonova, M. and Agrawal, S.

- (2020). Ceasa meets machine learning: A constant elasticity earth similarity approach to habitability and classification of exoplanets, *Astronomy and Computing* **30**: 100335.  
**URL:** <https://www.sciencedirect.com/science/article/pii/S2213133719300319>
- Bing, H. (2017). Statistical Machine Translation Algorithm Based on Improved Neural Network, *2017 International Conference on Robots Intelligent System (ICRIS)*, pp. 294–297.
- Bu, Y., Zeng, J., Lei, Z. and Yi, Z. (2019). Searching for Hot Subdwarf Stars from the LAMOST Spectra. III. Classification of Hot Subdwarf Stars in the Fourth Data Release of LAMOST Using a Deep Learning Method, *The Astrophysical Journal* **886**(2): 128. Publisher: American Astronomical Society.  
**URL:** <https://doi.org/10.3847/1538-4357/ab4c47>
- Bugueño, M., Mena, F. and Araya, M. (2018). Refining Exoplanet Detection Using Supervised Learning and Feature Engineering, *2018 XLIV Latin American Computer Conference (CLEI)*, pp. 278–287.
- Bugueño, M., Molina, G., Mena, F., Olivares, P. and Araya, M. (2021). Harnessing the power of CNNs for unevenly-sampled light-curves using Markov Transition Field, *Astronomy and Computing* **35**: 100461.  
**URL:** <https://www.sciencedirect.com/science/article/pii/S2213133721000159>
- Carado, B. and Knuth, K. H. (2020). Modeling transit and reflected light curves for non-spherical exoplanets, *Astronomy and Computing* **32**: 100406.  
**URL:** <https://www.sciencedirect.com/science/article/pii/S2213133720300603>
- Chaushev, A., Raynard, L., Goad, M. R., Eigmüller, P., Armstrong, D. J., Briegal, J. T., Burleigh, M. R., Casewell, S. L., Gill, S., Jenkins, J. S., Nielsen, L. D., Watson, C. A., West, R. G., Wheatley, P. J., Udry, S. and Vines, J. I. (2019). Classifying exoplanet candidates with convolutional neural networks: application to the Next Generation Transit Survey, *Monthly Notices of the Royal Astronomical Society* **488**(4): 5232–5250.  
**URL:** <https://doi.org/10.1093/mnras/stz2058>
- Chintarungruangchai, P. and Jiang, I.-G. (2019). Detecting Exoplanet Transits through Machine-learning Techniques with Convolutional Neural Networks, *Publications of the Astronomical Society of the Pacific* **131**(1000): 064502. Publisher: IOP Publishing.  
**URL:** <https://doi.org/10.1088/1538-3873/ab13d3>
- de Almeida, L., Anthony, F., Carolina Mattiuci, A., Castro, M., Soares da Costa, J., Samadi, R. and do Nascimento, J. D. (2020). Optimized cutting off transit algorithm to study stellar rotation from PLATO mission light curves, *Experimental Astronomy* **50**(1): 73–90.  
**URL:** <https://doi.org/10.1007/s10686-020-09664-x>
- de Franciscis, S., Pascual-Granado, J., Suárez, J. C., García Hernández, A., Garrido, R., Lares-Martiz, M. and Rodón, J. R. (2019). A fractal analysis application of the pre-whitening technique to Scuti stars time series, *Monthly Notices of the Royal Astronomical Society* **487**(3): 4457–4463.  
**URL:** <https://doi.org/10.1093/mnras/stz1571>

- Ferrari, A., Tourneret, J. and Schmider, F.-X. (1999). Detection of extra solar planets using parametric modeling, *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, Vol. 3, pp. 1369–1372 vol.3. ISSN: 1520-6149.
- Hinners, T. A., Tat, K. and Thorp, R. (2018). Machine Learning Techniques for Stellar Light Curve Classification, *The Astronomical Journal* **156**(1): 7. Publisher: American Astronomical Society.  
**URL:** <https://doi.org/10.3847/1538-3881/aac16d>
- Huo, J. (2018). Dynamic Time Warping and FFT: A Data Preprocessing Method for Electrical Load Forecasting, *International Journal of Advanced Computer Science and Applications* **9**(2).  
**URL:** <http://thesai.org/Publications/ViewPaper?Volume=9Issue=2Code=ijacsaSerialNo=1>
- Jagtap, R., Inamdar, U., Dere, S., Fatima, M. and Shardoor, N. B. (2021). Habitability of Exoplanets using Deep Learning, *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–6.
- Kalaei, M. J. and Hasanzadeh, A. (2019). A time series analysis of light curve of R Scuti star from 1970 to 2017, *New Astronomy* **70**: 57–63.  
**URL:** <https://www.sciencedirect.com/science/article/pii/S138410761930020X>
- Kirichenko, L., Radivilova, T., Bulakh, V., Zinchenko, P. and Saif Alghawli, A. (2020). Two Approaches to Machine Learning Classification of Time Series Based on Recurrence Plots, *2020 IEEE Third International Conference on Data Stream Mining Processing (DSMP)*, pp. 84–89.
- Linares, R., Furfaro, R. and Reddy, V. (2020). Space Objects Classification via Light-Curve Measurements Using Deep Convolutional Neural Networks, *The Journal of the Astronautical Sciences* **67**(3): 1063–1091.  
**URL:** <http://link.springer.com/10.1007/s40295-019-00208-w>
- Ozkok, F. O. and Celik, M. (2021). Convolutional neural network analysis of recurrence plots for high resolution melting classification, *Computer Methods and Programs in Biomedicine* **207**: 106139.  
**URL:** <https://www.sciencedirect.com/science/article/pii/S0169260721002145>
- Santos, L. A. d., Bourrier, V., Ehrenreich, D. and Kameda, S. (2019). Observability of hydrogen-rich exospheres in Earth-like exoplanets, *Astronomy & Astrophysics* **622**: A46. Publisher: EDP Sciences.  
**URL:** <https://www.aanda.org/articles/aa/abs/2019/02/aa33392-18/aa33392-18.html>
- Sartoli, S., Wei, Y. and Hampton, S. (2020). Malware Classification using Recurrence Plots and Deep Neural Network, *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 901–906.
- Shallue, C. J. and Vanderburg, A. (2018). Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90, *The Astronomical Journal* **155**(2): 94. Publisher: American Astronomical Society.  
**URL:** <https://doi.org/10.3847/1538-3881/aa9e09>



Sluijs, L. v., Mooij, E. d., Kenworthy, M., Celeste, M., Hooton, M. J., Mamajek, E. E., Sipócz, B., Snellen, I. A. G., Ridden-Harper, A. R. and Wilson, P. A. (2019). Spectroscopic transit search: a self-calibrating method for detecting planets around bright stars, *Astronomy & Astrophysics* **626**: A97. Publisher: EDP Sciences.

**URL:** <https://www.aanda.org/articles/aa/abs/2019/06/aa35066-19/aa35066-19.html>

Southworth, J., Bohn, A. J., Kenworthy, M. A., Ginski, C. and Mancini, L. (2020). A multiplicity study of transiting exoplanet host stars - II. Revised properties of transiting planetary systems with companions, *Astronomy & Astrophysics* **635**: A74. Publisher: EDP Sciences.

**URL:** <https://www.aanda.org/articles/aa/abs/2020/03/aa37334-19/aa37334-19.html>

Woodward, D., Stevens, E. and Linstead, E. (2019a). Generating transit light curves with variational autoencoders, *2019 IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT)*, pp. 24–32.

Woodward, D., Stevens, E. and Linstead, E. (2019b). Generating Transit Light Curves with Variational Autoencoders, *2019 IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT)*, pp. 24–32.