Predicting the Presence of Exoplanets in Star- Systems using Random Forest Classifier

Digvijay Patil
Computer Engineering and Technology
Dr. Vishwanath Karad MIT World Peace Universit,
Pune, India

Shraavya R. Srinivasarao,
Computer Engineering and Technology
Dr. Vishwanath Karad MIT World Peace University
Pune, India

Rajat Puri
Computer Engineering and Technology
Dr. Vishwanath Karad MIT World Peace University
Pune, India

***Abstract*—The hunt for exoplanets has been a focus in space exploration in order to discover habitable planets and star systems. NASA's Kepler (K2) space telescope collected periodic measurements of solar flux (light intensity) that were transmitted back to Earth, and the data was cleaned by NASA. This information can be used to draw conclusions about the potential presence of exoplanets in star systems of our galaxy. The Transit Method is one such method for detecting exoplanets, in which the difference in solar flux at different time intervals is observed; if the difference is large, an exoplanet will exist. The dataset consists of many observations having no exoplanets and very few observations with exoplanets. This imbalance in the data was overcome using the SMOTE (Synthetic Minority Over-sampling TEchnique) which increases the number of minority class variables while keeping the majority class variables constant. The prediction on the dataset was done using the Random Forest Classifier, an Ensemble based Machine Learning Algorithm which utilizes the dataset to the fullest. The composite model of SMOTE and Random Forest Classifier gave an accuracy of about 99% with optimal bias-variance trade-off. These findings can be used to intelligently screen out potential exoplanets, allowing us to focus our hunt for further research.**

***Keywords*—Exoplanets, Kepler, SMOTE, Random Forest, Solar Flux**

I. INTRODUCTION

The hunt for exoplanets has always been one of the primary focus of space research. The need to find a habitable planet is growing as the Earth warms. An Exoplanet is any planet outside of our solar system. Most orbit other stars, but free-floating exoplanets, called rogue planets, orbit the galactic centre and are untethered to any star[1]. The Mikulski Archive for Space Telescopes (MAST) make the Kepler's pixel and flux measurements publicly available. The Kepler has been observing close to 156,000 stars at 29.4 minutes interval as Long Cadence (LC) targets for primary purpose of detecting transiting planets[2]. The transit method relies on taking the light flux of the target star and comparing these flux values to other stars in the same patch of sky. When a planet passes in front of star the observed brightness dims down. This dimming depends on size of the planet. This results in a dip in observed flux values [3].

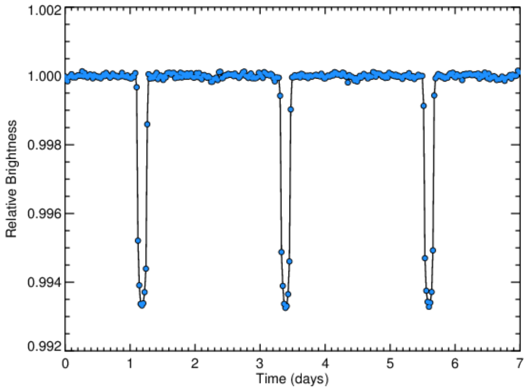

Fig. 1, Relative brightness vs Time

Several algorithms have been developed to effectively identify exoplanets. Machine Learning is one such avenue that has been used to determine whether an object of interest is a verified exoplanet or a false positive[4]. This study aims at understanding the effectiveness of Random Forest algorithm as a classification algorithm for classifying if a particular star can host an exoplanet or not. We use Precision, Recall, F1 scores, Confusion matrix, Area Under ROC Curve (AUC) and receiver operating characteristic curve (ROC) to evaluate the performance of the Machine Learning Model. This study uses the Kepler labelled time-series data for exoplanet hunting in deep space. This dataset contains two files, exoTrain and exoTest. The training set consists of 5087 observations and 3198 columns, column 1 is the labelled vector and column 2 to 3198 are flux values over time. The test set contains 570 observations and 3198 features[5].

II. BACKGROUND

An exoplanet is a planet that orbits another star in our galaxy outside of our solar system. Most exoplanets detected so far are concentrated in a tiny area of our galaxy, the Milky Way. A planet blocks some of the starlight as it passes directly between an observer and the star it orbits. The light of that star dims for a short time. It is just a small difference, but it's enough to alert astronomers to the existence of an exoplanet orbiting a distant star. This is known as the transit method.[6] The age, distribution, and composition of stars reveal information about the galaxy's origin, stability, and evolution. The creation and delivery of heavy elements including carbon, nitrogen, and oxygen are regulated by stars. One of the most profound questions of all time is whether life persists beyond Earth. Whether the answer indicates a cosmos rich in life, a universe where life is scarce and delicate, or even a universe where there is no trace of alien life, the answer stands to change us forever.

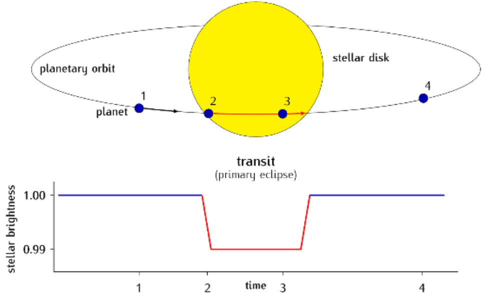

Fig. 2, Stellar brightness vs time: Explaining the Transit Method

The ultimate aim of this research is to discover undeniable indicators of contemporary life. Transmission spectroscopy, a method that analyses light fired by a star into the atmosphere of a distant planet, produces an effect that looks like a barcode. The light spectrum slices that are absent reveal the components are found in the alien environment. One sequence of black gaps could mean methane, while another could indicate oxygen. The coexistence of these make a strong case for existence of life. We may also read a barcode that indicates the combustion of hydrocarbons, or smog.

The importance of the study of exoplanets helps us understand the basics of existence of life in other planetary systems that are millions of light years away from our star system. Determining whether a particular star has planets that are revolving around it is the very first filter which eliminates star systems which have no planets.

Study of exoplanets helps us to get a basic understanding about the elements, hydrocarbons and possible alien life in our galaxy. This research aims to find solution to filter out the stars that don't inhibit planets in their system.

III. DATA ANALYSIS & DATA PRE-PROCESSING:

A supervised Machine Learning model's generalisation efficiency is greatly influenced by data pre-processing. J. Sola and J. Sevilla found that prior to model training, adequate normalisation of input data provided clear benefits: The number of estimation errors was decreased by a factor of five to ten, and the measurement time was reduced by an order of magnitude.

A. Data Source:

The data provided here has been cleaned and is obtained from NASA's Kepler space telescope observations. Verified exoplanets from other efforts were also used to increase the number of exoplanet-stars in the dataset. NASA open-sources the original Kepler Mission data and it is hosted at the Mikulski Archive[7]. After being beamed down to Earth, NASA applies de-noising algorithms to remove artefacts generated by the telescope.

B. Data Analysis:

The dataset contains the solar fluxes of several stars and planetary systems at several time intervals. The data describes the change in flux, also known as light intensity, of several thousand stars. Each star has a binary label of 2 or 1. 2 indicated that that the star is confirmed to have at least one exoplanet in orbit; some observations are in fact multi-planet systems. The planets themselves do not emit light, but the stars that they orbit do. If said star is observed over a period, there may be a regular 'dimming' of the flux. This is proof that there may be an orbiting body around the star. Such a star may be termed as a 'candidate system'. Further investigation of our candidate scheme, such as by a satellite capturing light at a different wavelength, could strengthen the assumption that the candidate can be 'confirmed.'

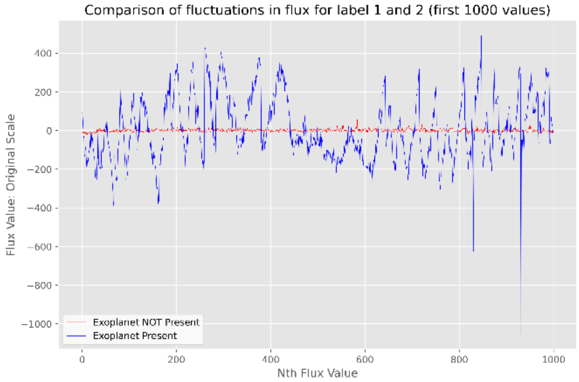

Fig. 3, Analysis of Flux values of two observations

The training dataset contains 5087 observations and flux values at 3197 time intervals for each observation. In those many observations only 37 observations have the label 2 indicating that the star has an exoplanet in the orbit of the parent star. This implies that the dataset is highly imbalanced as 0.7% of the available observations are of Class A (Label 2) and 99.3% of Class B (Label 1). Thus, a method called SMOTE is used to balance the data and cover the difference between the number of data points of 2 classes.

C. Workflow of SMOTE:

Synthetic Minority Over – Sampling TEchnique is a statistical technique for increasing the number of cases in your dataset in a balanced way. The module works by generating new instances from existing minority cases that you supply as input. This implementation of SMOTE does not change the number of majority cases. The new instances are not just copies of existing minority cases; instead, the algorithm takes samples of the *feature space* for each target class and its nearest neighbors and creates new examples that combine features of the target case with features of its neighbors. This approach increases the features available to each class and makes the samples more general.[8]

In our dataset, the SMOTE is applied after a series of basic data preprocessing steps.

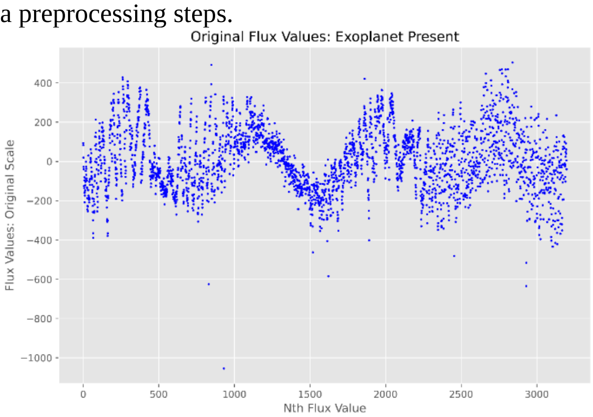

Fig. 4, Plot of Original Flux Values at Several Time Intervals: Exoplanet Present

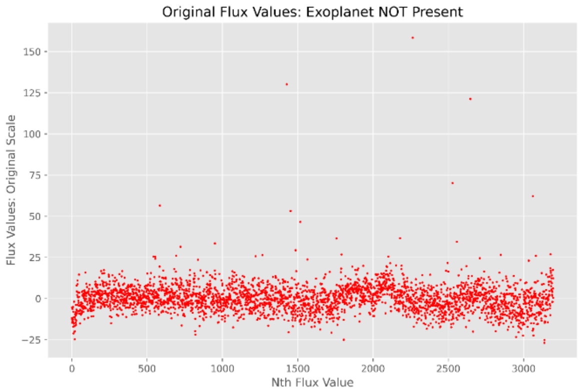

Fig. 5, Plot of Original Flux Values at Several Time Intervals: Exoplanet NOT Present

The following preprocessing steps are performed before SMOTE is applied.

D. Fourier Transform:

The first function that is used is the Fourier transform on the dataset. The dataset, as we know, includes flux values at various time intervals. As a result, we transform the time domain values of superimposed sinusoidal waves into frequency domain values. The Fourier representation has the advantage of capturing data distribution nonlinearities without requiring the definition of a kernel function. It also allows for a probabilistic interpretation of classification, unlike support vector machines. It can also deal with groups that overlap. Unlike logistic regression, Fourier representation does not require feature engineering. In general, its computational performance is also very well for large data sets and in contrast to other algorithms, the typical overfitting problem is not seen. The algorithm's ability to perform multiclass classification with overlapped classes and extremely nonlinear class distributions is demonstrated.[9] The following figure illustrates the conversion from given data values to Fourier data values.

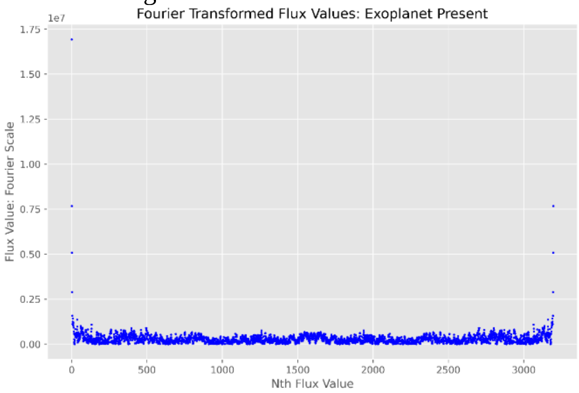

Fig. 6, Plot of Fourier Transformed Flux Values at Several Time Intervals: Exoplanet NOT Present

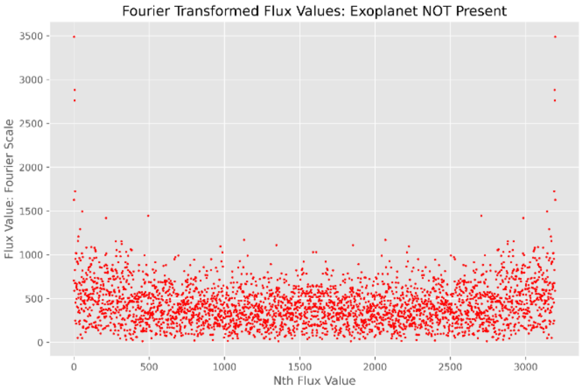

Fig. 7, Plot of Fourier Transformed Flux Values at Several Time Intervals: NOT Exoplanet Present

E. Savgol Filter:

It is analyzed by looking at a single observation from the dataset and plotting the flux values at several time intervals that the points are scattered across the highest and the lowest values. This gives a number of outliers which will affect the training of our Machine Learning model. Thus, the dataset must be processed so that the transition from one point to another point is smooth. We use the Savgol (Savitzky–Golay) filter. It is a digital filter that can be applied to a set of digital data points for the purpose of smoothing the data, that is, to increase the precision of the data without distorting the signal tendency. The idea behind SG smoothing is quite simple. For each data point in the spectrum, the SG algorithm will:

- Select a window (say, five points) around that point
- Fit a polynomial to the points in the selected window
- Replace the data point in question with the corresponding value of the fitted polynomial.[10]

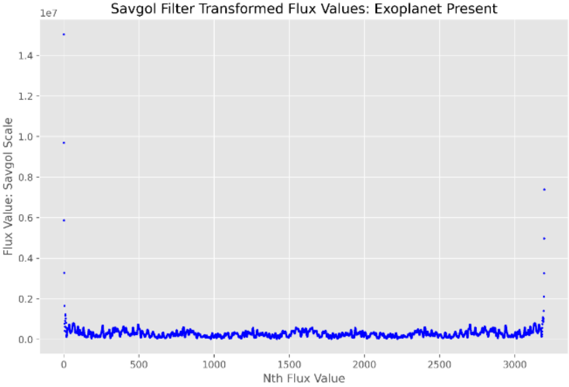

Fig. 8, Plot of Savgol Filtered Flux Values at Several Time Intervals: Exoplanet Present

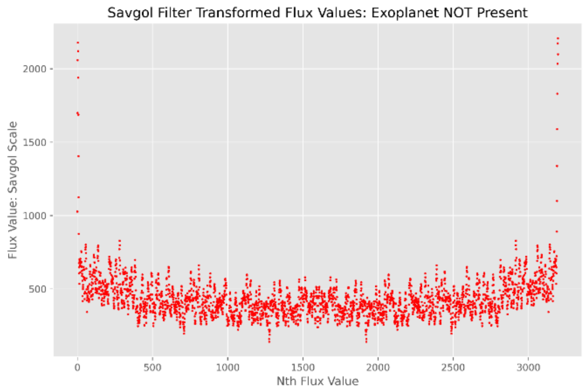

Fig. 9, Plot of Savgol Filtered Flux Values at Several Time Intervals: Exoplanet NOT Present

Applying Savgol filter to the output of Fourier function provides us with a smoothened data that is also comparatively free from outliers.

F. *Normalize:*

Normalization is a technique often applied as part of data preparation for machine learning. Normalization is the process of converting the values of numeric columns in a dataset to a common scale without distorting the ranges of values or losing details. Some algorithms require normalization to correctly model the data. When attempting to combine the values as features during modelling, the large difference in size of the numbers can cause problems. Normalization solves these issues through generating new values that preserve the source data's general distribution and ratios while holding values within a scale that is applied to all numeric columns throughout the model.

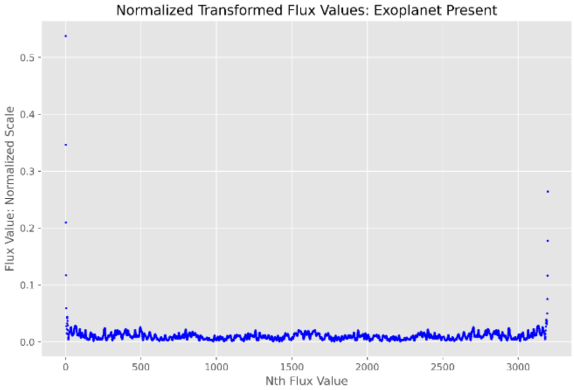

Fig. 10, Plot of Normalized Flux Values at Several Time Intervals: Exoplanet Present

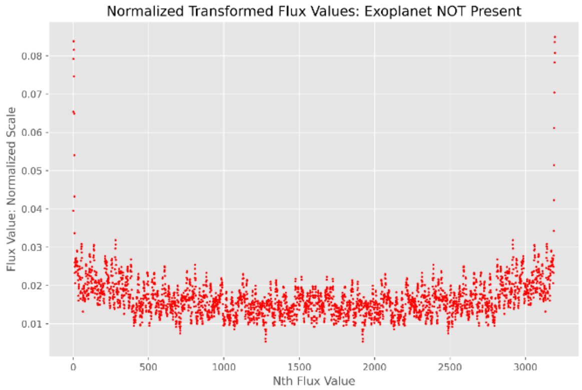

Fig. 11, Plot of Normalized Flux Values at Several Time Intervals: Exoplanet NOT Present

The output of Savgol filter is normalized to get the high range values obtained from it to a range that works best with the machine learning algorithm.

G. *Robust Filter:*

Even after the application of Savgol Filter, the data still contains outliers that are now pulled down to the normalized range. These outliers need to be treated before the application of SMOTE. It uses outlier-resistant statistics to scale functions. This approach eliminates the median and scales the data between the first and third quartiles, or between the 25th and 75th quantile range. This range is also called Interquartile range. The median and interquartile range are then stored so that the transform method can be applied to future results. If the dataset contains outliers, the median and interquartile range produce better results and outperform the sample mean and variance.[11]

Thus, the output of robust filter can be visualized in the following figure.

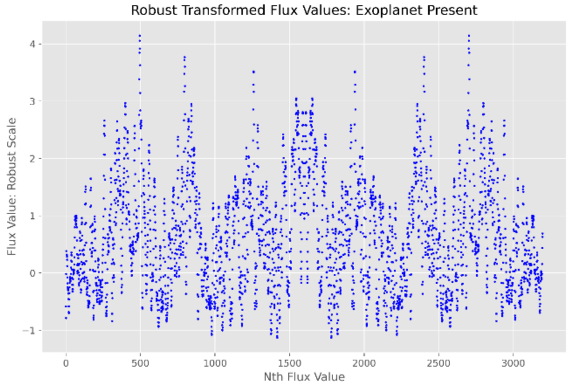

Fig. 12, Plot of Robust Transformed Flux Values at Several Time Intervals: Exoplanet Present

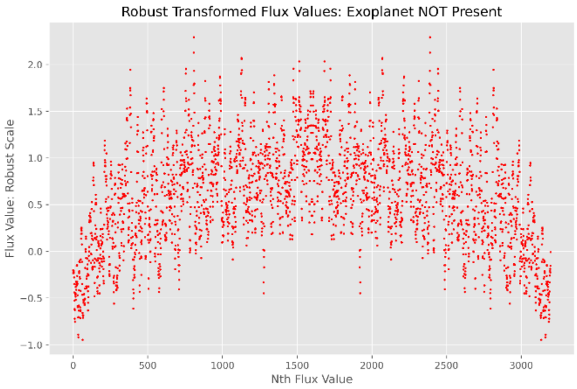

Fig. 13, Plot of Robust Transformed Flux Values at Several Time Intervals: Exoplanet NOT Present

H. SMOTE:

As stated in the above section, the dataset we have is highly imbalanced and we need to balance the data so that the model doesn't cause a bias and variance problem. This problem is solved using the application of SMOTE on the dataset. The SMOTE function will loop through the given dataset and find the existing and real minority class instances. At each iteration of the loop, one of the K closest minority class neighbors is chosen and a minority class instance is introduced and synthesized in between the minority instance and that neighbor.

Thus, the SMOTE is the final pre-processing step that is used before the data is fed to the classification algorithm. The results from SMOTE are illustrated in the given figure.

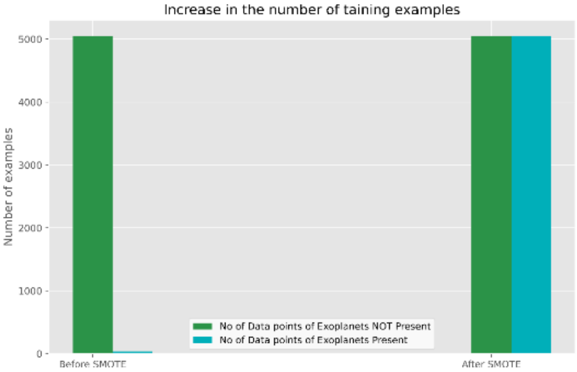

Fig. 14, Dataset after SMOTE Transformation

IV. METHODOLOGY

The processed data is to be classified between class label 1 and class label 2. This is a binary classification and can be achieved with several Machine Learning algorithms. The pre-processing with the inclusion of SMOTE helped in generating a balanced data. The algorithm that is used for classification is Random Forest Classifier.

Random Forest is an ensemble learner, which produces many classifiers and aggregates their output. To evaluate the split, Random Forest will construct several classification and regression (CART) trees, each of which will be trained on a bootstrap sample of the original training data and will search across a randomly selected subset of input variables[12]. This method involves introducing a random variable into the construction of successive decision trees. According to the findings by Khalilia Chakraborty and Popescu, the test set misclassification percent of Random Forest was found to be 50% less than that of Decision Tree[13]. Random Forest can handle high dimensional data and uses many trees in the ensemble.

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. So, a Random Forest Classifier uses several Decision Trees and uses a method called as ensemble learning which creates a voting average pool for predicting the final class of the input data.

The decision to make strategic splits has a significant impact on a tree's accuracy. The decision criteria for classification and regression trees are different. To determine whether to divide a node into two or more sub-nodes, decision trees employ a variety of algorithms [14]. The homogeneity of the resulting sub-nodes improves with the construction of sub-nodes. In other words, the purity of the node improves as the goal variable increases. The decision tree divides the nodes into sub-nodes based on all available variables, then chooses the split that produces the most homogeneous sub-nodes. The reason why Random Forest is preferred over other algorithms is:

- It produces a highly accurate classifier.
- Learning is faster.
- It can handle thousands of input variables without variable deletion.
- It computes the proximities between pairs of cases that can be used in clustering or locating outliers.
- It calculates the significance of classification variables.
- It has a method for balancing errors in unbalanced data called weighted random forest (WRF).
- It has an effective method for predicting missing data.[13]

In our model the Random Forest classifier takes all the 3197 time instances at once and generates decision trees based on the splits identified by the classifier. The classifier generates 100 Decision Trees and a voting approach outputs the final class of the input data. The depth of the trees is not specified as nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples which is default 2 as per the sklearn documentation.

The following figure illustrates a typical Random Forest Architecture.

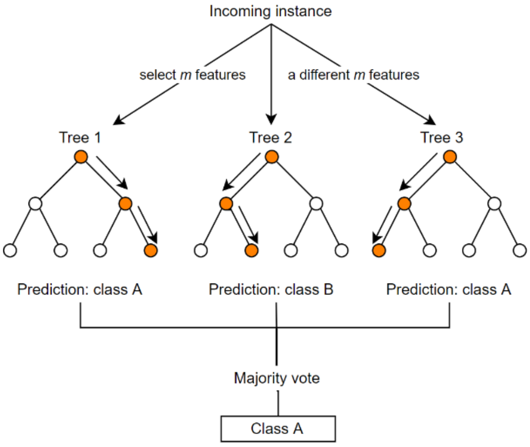

Fig. 15, A typical random forest architecture

V. RESULTS:

The accuracy of the model is analysed by a number of methods. The following are the accuracy metrics that are used to analyse the performance of the model and their interpretation:

A. *Classification Report*:

The classification report gives a clear idea about how many data points are classified correctly and how many are misclassified.

- Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP/TP+FP

The model gives a precision value of 1.0 indicating that maximum class B (Label 1) values are predicted correctly.

- Recall: Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

Recall = TP/TP+FN

The model gives a recall values of 1.0 indicating that the model correctly predicted maximum values of Class B (Label 1) while not being biased towards the class. Thus, indirectly implying that the predictions of Class A (Label 2) are also maximum.

- F1 Scores: The accuracy of the classification is calculated using F1 score. The F1 Score is calculated as:

2 * ((precision * recall) / (precision + recall)).

It is also called the F Score or the F Measure. In other words, the F1 score conveys the balance of precision and recall. The model gives a F1 Score of **0.999045** indicating that both classes were predicted rightly by maximum number of input data values.

The following table summarizes the classification report for both the classes.

TABLE I. CLASSIFICATION REPORT AND ACCURACY SCORES

Classification Report	*Precision*	*Recall*	*F1-Score*	*Support*
1	1.00	1.00	1.00	2028
2	1.00	1.00	1.00	1572
Accuracy			1.00	3600
Macro Average	1.00	1.00	1.00	3600
Weighted Average	1.00	1.00	1.00	3600

[a] *Classification Report of Predicted Values*

B. *Confusion Matrix*:

The confusion matrix gives the exact number of rightly predicted values and wrongly predicted values. Out of the total 3600 values in the test dataset, which contained, after the application of SMOTE, 2028 Class B (Label 1) values and 1572 Class A (Label 2) values, the model rightly classified 2027 values from Class B and 1570 as Class A. Thus only 3 values from the total test set were misclassified by the Random Forest Classifier.

The following figure illustrates the confusion matrix.

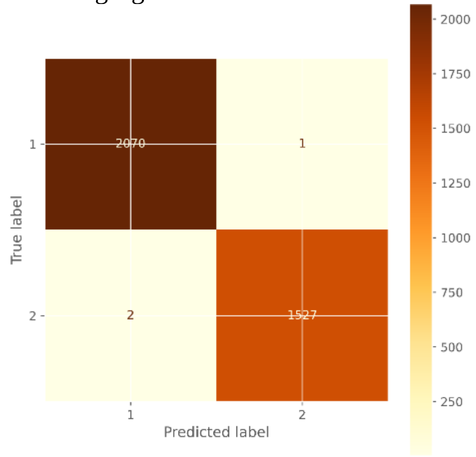

Fig. 16, Confusion matrix for Random Forest Classifier

C. *True Positives vs False Positives:*

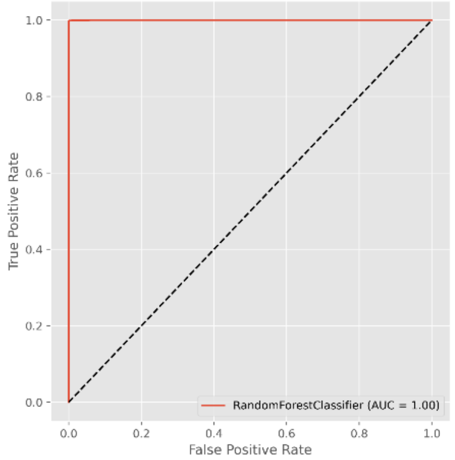

Fig. 17, True Positives rate vs False Positives

D. *True Positives vs Recall Graph:*

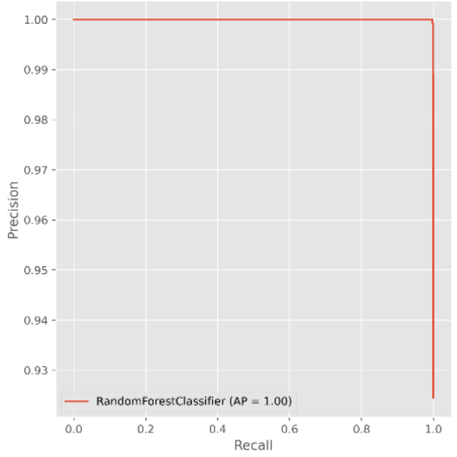

Fig. 18, Precision vs Recall

The overall efficiency of the model increased because of the inclusion of SMOTE in the pre-processing steps.

VI. FUTURE STEPS:

The research can be extended to find more accurately the presence of exoplanets in the star systems of the galaxy. The dataset that was used currently was obtained from the Kepler (K2) space telescope. Further this data can be increased and added to the archives of data collected from the TESS – Transiting Exoplanet Search Satellite. The TESS and K2 combined can provide a much larger dataset that contains not only the solar flux from the nearby stars but also their images captured by the satellites. The images with the computer vision techniques and use of Convolutional Neural Networks can be used to identify the precision of the results obtained from our current model.

Further a study from NASAs TESS data revealed the internal structures of the stars, which could aid in the understanding of what's happening in billions of stars across the universe by analysing the patterns of pulsating stars.[15] This helps us in narrowing down the stars that have exoplanets by identifying their patterns and applying similar techniques to each of the star present in the observable universe to understand whether the star has an exoplanet or not. This reduces are analysis time and complexity where solar flux is no longer the main candidate of prediction but wavelengths of flux captured from the satellites.

The most important breakthrough that the research is aiming towards is the analysis of compounds and elements in the environment of the detected exoplanets and their stars. Earth-like exoplanets are between 10 million and 10 billion times fainter than the stars they orbit, depending on whether they are observed at mid-infrared or visible wavelengths. The presence of water, methane, and oxygen on a planet, revealed by reflected starlight, is therefore masked by the star's overwhelming glare.[16]

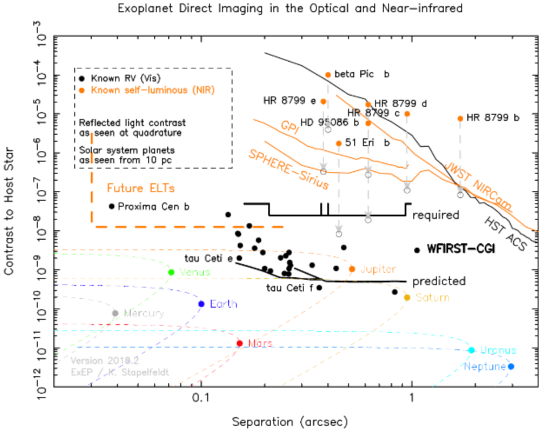

Fig. 19, Contrast (ratio of planet brightness to host star brightness) versus apparent angular separation. The filled orange circles indicate the direct imaging of young, self-luminous planets imaged in the near-infrared by ground-based telescopes. Contrasts for the planets of the Solar System are for analogous planets placed 10 pc away. The solid black dots are contrast estimates of measured radial velocity planets, including Proxima Cen b. The orange curves show measured performance of ground-based coronagraphs. The GPI curve shows typical performance, while the SPHERE curve shows the best achieved performance to-date on Sirius. Achieved performance with HST/ACS coronagraphic masks, and the predicted performance of JWST/NIRCam masks are also shown. The predicted and required performance at 565 nm for the WFIRST coronagraph instrument (CGI) is shown as solid black curves. The "predicted" curves extending from 0.13" to 0.4" is based on performance achieved in a testbed. From 0.4" to 1", performance is based on a coronagraph mask designed to maximize outer working angle. For consistency, the planets discovered in the near-infrared are shown with vertical arrows pointing to the predicted contrast ratios at visible wavelengths (WFIRST-CGI is expected to conduct science between 442 and 980 nm).

The aim of this research is not only to identify exoplanets, but also to determine if the planet is habitable and a possible candidate for further exploration. Using data archives from K2, TESS, and NASA's Exoplanet Program, the techniques of Starlight Suppression, Wave front Control with better Detection Sensitivity, and Angular Resolution can be used to address the current challenges of masked glare, increased intensity, and denoising the data that is usable for study, thus leading the way to the discovery of habitable exoplanets

VII. CONCLUSION:

With new and advanced telescopes, data in astronomy are growing at a fast pace. Conventional methods that involve human judgements are not efficient and prone to variability depending on the investigating expert. In this study, we propose a novel exoplanet detection method based on Random Forest Classification model machine learning.

The Random Forest Classifier Model, when paired with the pre-processing steps of Fourier Transforms and SMOTE, accurately predicted whether a star system has exoplanets or not, with a prediction accuracy of 99.00 percent. The pre-processing steps account for a major contribution in the classification and prediction as the data that is available is highly imbalanced and biased towards the class that states that the star doesn't have any exoplanet or there is less possibility of the star having a planet in its proximity.

This research can aid scientists in further understanding planetary systems and will serve as a second confirmation stage in determining whether a star has exoplanets. While traditional methods of detection are accurate, the aim of this research is not to restrict itself to planet detection; rather, in the distant future, it hopes to detect potential hydrocarbons and elements in the planets' environments, confirming the presence of life or the likelihood of another habitable planet in the galaxy.

REFERENCES

[1] "Overview | What is an Exoplanet?" *Exoplanet Exploration: Planets Beyond our Solar System*. https://exoplanets.nasa.gov/what-is-an-exoplanet/overview (accessed Apr. 21, 2021).

[2] J. M. Jenkins *et al.*, "INITIAL CHARACTERISTICS OF *KEPLER* LONG CADENCE DATA FOR DETECTING TRANSITING PLANETS," *Astrophys. J.*, vol. 713, no. 2, pp. L120–L125, Apr. 2010, doi: 10.1088/2041-8205/713/2/L120.

[3] "Transit Light Curve Tutorial." https://lweb.cfa.harvard.edu/~avanderb/tutorial/tutorial.html (accessed Apr. 21, 2021).

[4] "View of Exoplanet Hunting in Deep Space with Machine Learning." https://www.journals.resaim.com/ijresm/article/view/323/298 (accessed Apr. 21, 2021).

[5] "Exoplanet Hunting in Deep Space." https://kaggle.com/keplersmachines/kepler-labelled-time-series-data (accessed Apr. 21, 2021).

[6] "In Depth | What is an Exoplanet?" *Exoplanet Exploration: Planets Beyond our Solar System*. https://exoplanets.nasa.gov/what-is-an-exoplanet/in-depth (accessed Apr. 27, 2021).

[7] "K2," *MAST*. http://archive1.stsci.edu/home/missions-and-data/k2 (accessed Apr. 27, 2021).

[8] xiaoharper, "ML Studio (classic): SMOTE - Azure." https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote (accessed Apr. 27, 2021).

[9] S. Mehrabkhani, "Fourier Transform Approach to Machine Learning III: Fourier Classification," *ArXiv200106081 Cs Stat*, Mar. 2020, Accessed: Apr. 27, 2021. [Online]. Available: http://arxiv.org/abs/2001.06081.

[10] D. Pelliccia, "Savitzky–Golay smoothing method • NIRPY Research," Oct. 05, 2019. https://nirpyresearch.com/savitzky-golay-smoothing-method/ (accessed Apr. 27, 2021).

[11] "StandardScaler, MinMaxScaler and RobustScaler techniques - ML," *GeeksforGeeks*, Jul. 15, 2020. https://www.geeksforgeeks.org/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/ (accessed Apr. 27, 2021).

[12] R. Puri and D. Patil, "Comparative Study of Machine Learning Algorithms on Binary Dataset," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 137–147, Mar. 2021, doi: 10.48175/IJARSCT-887.

[13] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, p. 51, Dec. 2011, doi: 10.1186/1472-6947-11-51.

[14] “Decision Tree Algorithm, Explained,” *KDnuggets*. https://www.kdnuggets.com/decision-tree-algorithm-explained.html/ (accessed Apr. 27, 2021).

[15] A. S. CNN, “The ‘beating hearts’ of these pulsating stars create music to astronomers’ ears,” *CNN*. https://www.cnn.com/2020/05/15/world/pulsating-stars-delta-scuti-scn-trnd/index.html (accessed Apr. 28, 2021).

[16] “Exoplanet Program: Technology Overview,” *Exoplanet Exploration: Planets Beyond our Solar System*. https://exoplanets.nasa.gov/exep/technology/technology-overview (accessed Apr. 28, 2021).