Exoplanet Detection using Feature Engineering with Ensemble Learning

G. Venkata Sai Rakesh
School of CSE
VIT-AP University
Amaravati, India
rakigollamandala@gmail.com

M. Jahnavi Bhuvana Chandrika
School of CSE
VIT-AP University
Amaravati, India
mjahnavi263@gmail.com

Ch. Venkata Rami Reddy
School of CSE,
VIT-AP University,
Amaravati, India
chvrr58@gmail.com

Muvva Suneetha
Independent Researcher
Amaravati, India
sunee.muvva@gmail.com

***Abstract*— In this work, a novel technique for detecting extra-terrestrial bodies using the transit method, with the aim of improving traditional algorithmic strategies in astronomy through machine learning algorithms. The potential for this approach to revolutionize the field of exoplanet detection and unlock new insights into the universe is highlighted. An ensemble learning approach using majority voting technique is employed to detect Exoplanets. In this work, we have used six machine learning models Random Forest, Decision Tree, Support Vector Classifier, K-Nearest Neighbor and Multi-Layer Perceptron, for Exoplanet Detection. Based on accuracy, we combined Support vector Classifier, K-Nearest Neighbor, Random Forest and Multi-Layer Perceptron using Majority Voting Technique to predict whether the planet is exoplanet or not. Majority Voting Technique (MVT) of machine learning models has shown higher significance when compared to the individual machine learning model in terms of accuracy. The accuracy of different individual models is ranging from 92.59% to 99.88% and MVT produced 99.97% accuracy.**

Keywords—Transit Method, Extra-terrestrial, Ensemble Learning, Majority Voting Technique

I. INTRODUCTION

The Solar System is believed to have formed over a period of 4600 million years. The possibility of the presence of numerous stars in other solar systems has long been a subject of contemplation. In the last two decades, data from numerous exoplanetary systems have been collected, leading to the discovery that planetary systems do not form according to the well-ordered solar nebula model, but rather through chaotic processes. Small rocky planets tend to be tightly packed in interior orbits, while gas giant planets can move closer to their stars. Exoplanets are celestial bodies located outside our solar system that share similarities with Earth in terms of climate and atmosphere, making them potential sites for human habitation. The first exoplanet, known as PULSAR, was discovered in 1992 by two radio astronomers, leading to the awareness of the possibility of alien life. As of March 1, 2023, there are 5,332 verified exoplanets in 3,931 planetary systems. Astrophysicists have devised a number of methods for finding exoplanets, which have become a major research topic in the field. In order to better comprehend the possibility of extra-terrestrial life, exoplanet analysis looks at how they interact with their host star, as well as the surface chemistry and atmosphere. It is possible to detect and analyse exoplanet atmospheres, identify chemical signatures on their surfaces, and measure their physical properties by using observational methods like spectroscopy, photometry, and direct imaging. Our knowledge of the likelihood of life existing elsewhere in the universe has been greatly impacted by these findings. Exoplanets come in a variety of shapes and sizes, including hot Jupiter's, gas giant planets that orbit close to their host star, super-Earths, rocky planets with masses halfway between Earth and Neptune, and potentially habitable planets that orbit in their star's habitable zone, where conditions are favourable for the presence of liquid water. The classification of exoplanets is based on their mass, size, and separation from their star, and has been achieved through multiple observational techniques such as radial velocity measurements, transit photometry, and direct imaging.

The transit approach, one of the techniques for finding exoplanets, can be identified using machine learning algorithms. The Kepler space telescope, operated by NASA, has played a pivotal role in making this technique more widely used. Instead of directly observing exoplanets from afar, transits identify them by detecting the dimming of their host star's light as they pass in front of it. Light curves, which are graphs that display the quantity of light absorbed over time, can be used to see this dimming. The light curve's brightness will drop as the exoplanet gets closer to its host star. Figure 1. shows the transit light curve, which gauges the planet's thermal radiation at various wavelengths.

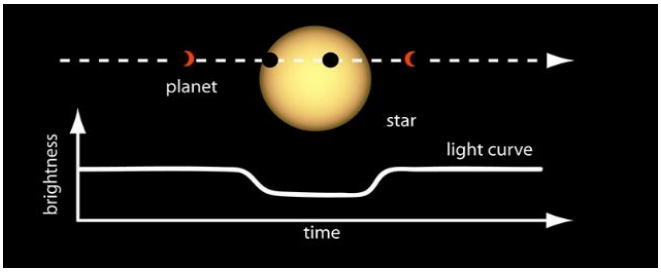

Fig. 1. Transit light curve.

II. RELATED WORK

In previous studies [1-9], deep learning and machine learning models have been commonly used for Exoplanet Detection. However, in our study, we opted for a machine learning approach. Although deep learning models are highly effective in analyzing complex data, machine learning models offer a higher degree of interpretability than deep learning models. Additionally, machine learning models are more data-efficient and are capable of providing comparable or even better performance than deep learning models with

less data. Using feature engineering in conjunction with machine learning models can be an effective approach as it enables the selection of the most suitable features.

Koray Aydogan[10] used Machine Learning with Data Augmentation for Exoplanet Detection. In this work they used noise augmentation and GAN's architecture and achieved accuracy of 97%. Abhishek Malik, et al [11] identify exoplanets using the transit method and the author proposes a method based on a tree-based classifier and the well-known machine learning tool lightgbm. Using the time-series analysis package TSFresh, they extrapolate 789 features from light curves. Each light curve's features are recorded in these traits. They trained and evaluated this technique using real Kepler and TESS data as well as synthetic data. It was found to be superior to traditional box least squares fitting in the assessment using synthetic data. The method can accurately and reliably identify a planet transit on Kepler data with an AUC of 94.8% and a Recall of 96%. The method can classify light curves with a 98% accuracy using the TESS data and can find planets with an 82% recall. "Machine Learning Pipeline for Exoplanet Classification was used by Sturrock GC et al[12]. SVM, KNN, and RF are three classification algorithms were used in order to determine the probability that an observation is an exoplanet. The Cumulative Kepler Object of Interest (KOI) database, which collects data for all Kepler Objects of Interest (KOI), was used to categorize the data, and the best machine learning model was selected. On the training set, the Random Forest Classifier had a cross-validated accuracy score of 98.96%, a precision of 99.55%, and a recall of 97.21%.

The convolutional neural network was used by Christopher J. Shallue1 and Andrew Vanderburg [13] to Identify exoplanets. Features are recovered by folding each flattened light curve in the TCE interval and clustering to obtain a 1D vector. The training and evaluation sets were chosen from the Q1-Q17 DR24 Autovetter Planet Candidate Catalog. They compared models based on linear logistic regression and fully connected neural networks and achieved 95% recall, 90% accuracy, and 96% precision. Automated Triage and Vetting of TESS Candidates" by Liang Yu et al [14] to Identify exoplanets with Deep Learning technique. The deep neural network was developed and tested using actual TESS data, and it has an average precision and accuracy of 97.0% and 97.4% for identifying transit-like signals from instrumental noise in triage mode.

Pattana Chintarungruangchai and Ing-Guey Jiang [15] present a CNN-based technique for detecting exoplanet transits. In order to create a collection of images for training, a method for 2D phase folding is suggested. With or without folding, they test the technique using five distinct kinds of deep learning models. All folding versions have accuracy levels above 98%. Without folding, models' precision can increase to about 85%.

III. PROPOSED WORK

A. Process framework

The process frame was given in figure 2. The loading of data and exploratory data analysis are the first two stages in the training of a machine learning model. Following data loading, it's critical to comprehend its characteristics and spot any tends that might prove helpful during model training. exploratory data analysis, which entails displaying the data and gleaning insights. To convert the data into a format suitable for model training, feature engineering is used. The Synthetic Minority Over-Sampling Technique (SMOTE), was used to deal the imbalance data. To improve the performance of the model, ensemble learning was used.

B. Data Pre-processing

We have completed data preparation through data normalisation. After ensuring that there were no missing data in the dataset, we moved on to a crucial pre-processing step: normalising the data. We used normalisation on both the testing and training datasets to guarantee that they received the same normalisation treatment, which is important for accurate and consistent model performance.

Overall, the goal of normalising the data in this manner is to ensure that the numerical characteristics in the datasets are on the same scale and follow a specific distribution. This can help to enhance the accuracy and performance of data-driven machine learning models.

C. Feature Engineering

Feature engineering is a crucial step in enhancing the performance of machine learning algorithms, involving the selection, transformation, and creation of input features based on the data. This process requires domain expertise and creativity to identify relevant features that can predict the target variable accurately.

a) Feature Creation: The method of creating new features from already-existing data is known as feature creation. For instance, the rolling mean function is an effective method to create new features by calculating the average values of time series data.

b) Feature Transformation: Feature transformation is another important technique that involves scaling numerical data to have a standard deviation of 1 and a mean of 0 using the standard scaler method. This process is performed independently on each feature of the data.

c) Feature Extraction: Principal Component Analysis (PCA) is a method employed for the purpose of reducing the dimensionality of a given dataset. The technique involves the identification of the essential features or components that account for a significant amount of the variance present in the data. By transforming the correlated features into uncorrelated principal components, the technique effectively reduces the number of dimensions, while retaining the majority of the original information. In this particular dataset, we have chosen to utilize 30 principal components, chosen based on the covariance matrix, to represent the significant variations in the original features. This reduction in dimensionality provides several benefits, including improved performance of machine learning models and the avoidance of overfitting.

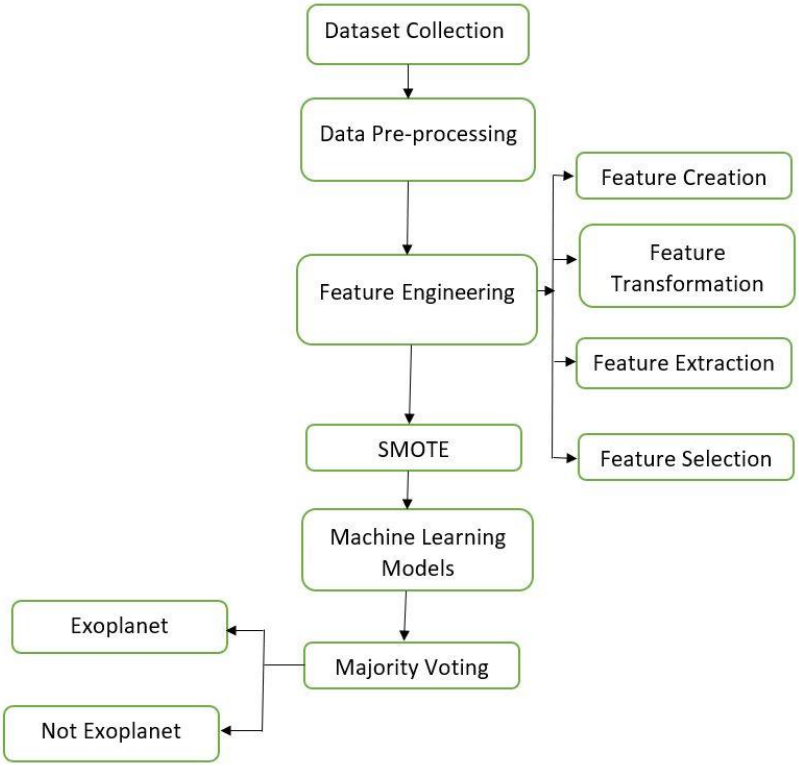

Fig. 2. Proposed framework

d) Feature Selection: The top k features that are most relevant to the target variable can be found by using feature selection, which is also very important. Based on each feature's individual scores and the relationship between it and the target variable, the best 5 features were chosen using the SelectKBest and f_classif techniques, respectively.

D. SMOTE

To address imbalanced datasets, a useful approach is SMOTE, a data augmentation technique that generates synthetic examples of the minority class by interpolating between existing ones. The newly created examples help to enhance the minority class's representation in the dataset, ultimately balancing it. SMOTE is typically employed when datasets are highly imbalanced, and the minority class is underrepresented. This is critical for preventing biases in the model's predictions, particularly in applications such as fraud detection, medical diagnosis, and image classification. SMOTE is a widely used technique that improves the model's performance and reduces bias in such situations.

E. Support Vector Machine

SVM is a machine learning algorithm used for classification tasks. SVM determines the optimal hyperplane that divides the data elements into various classes.

a) Linear Kernal: When the input data can be divided into distinct classes using a hyperplane and is thus linearly separable, the linear kernel is frequently employed in SVMs. The linear kernel is a suitable option in this situation because it is computationally effective and frequently yields decent results.

F. Random Forest

Classification and regression-related problems are resolved using Random Forest. As a consequence of the training process, which involves the construction of several decision trees, the mean class, or average class, of each tree is either categorised or regressed. Feature bagging, a potent and reliable technique, is used to choose randomly produced subsets of features for each tree, and bootstrap sampling, a technique to gather randomly generated samples from the training data.

G. KNN

KNN works by identifying the K data points in the training set that are closest to the input and generating the most frequent classification (classification) or mean value (regression) of the K closest neighbors as the prediction. Although it can be computationally expensive for big datasets, KNN is a simple and efficient algorithm.

H. Multi-Layer Perception

Multi-Layer Perceptron Classifier is a type of artificial neural network used for classification tasks. It consists of multiple layers of neurons, each connected to the next, and uses backpropagation to learn the weights and biases that map inputs to outputs. It is a popular and powerful algorithm for a wide range of classification tasks.

I. Decision Tree

An analysis of choices and their possible outcomes is done using a decision tree, a visual model. It is made up of a branching structure that begins with a single choice or query and develops into a number of potential outcomes or directions depending on various circumstances or variables. With each branch representing a different option or course of action, the model can be used to determine the likelihood of each result based on the incoming data and variables.

J. Majority Voting

Majority voting, which combines the output of numerous models to produce a singular forecast, is a popular technique in ensemble learning. The class label with the greatest frequency is the ensemble model's final output.

IV. Results And Discussions

We have used Kepler data set to detect the exo-planets. The value in the data set are noted by the Kepler spacecraft by using a photometer to continuously monitor the brightness of more than 150,000 stars in a specific region of the sky. This technique is known as transit photometry and it involves measuring the slight decrease in a star's brightness when a planet passes in front of it, blocking a small fraction of its light. The data in the dataset describes how the movement of thousands of stars has changed. Each star is given a binary designation of 2 or 1. The number 2 signifies that the star has at least one exoplanet that has been confirmed to be in its orbit; some discoveries are in reality multi-planet systems. There are 5087 rows and 3198 columns of events in the Train data. Column 2 contains the flow values over time, and Column 1 contains the name vector. There are 5050 non-exoplanet stars in the train data in addition to the 37 verified exoplanet stars. There are 3198 columns and 570 rows in the test data.

The study utilized unique methods that were considered to be more accurate than previous works on the subject. The results were not solely dependent on one algorithm but employed ensemble learning, which integrated multiple algorithms to obtain findings that were more precisely calculated. A combination of four machine learning algorithms - Random Forest, KNN, MLP, and SVC - was used to achieve an accuracy of 99.97%, surpassing the accuracy achieved by each technique used separately. The approach is considered to be highly effective, and the findings presented are expected to be of great interest. There are five confirmed exoplanet stars in the test results. In this work, we used accuracy, confusion matrix, precision, recall, and F1 score to evaluate our proposed models. Equations 1 to 4 shows the Accuracy, Precision, Recall and F1-score respectively.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (2)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (3)$$

$$\text{F1-Score} = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

TP = True Positive TN = True Negative

FP = False Positive FN = False Negative

Recall is a machine learning performance metric that calculates the proportion of true positives in comparison to all actual positives. When reducing false negatives is a top priority and the cost of false negatives is high, recall can be helpful. When the classes are unbalanced, the F1 score is a measurement that strikes a compromise between precision and recall. Table 1 shows the accuracy of proposed models. We used five machine learning models to determine the accuracy, and then used the majority voting technique to determine the accuracy by selecting the top four models that provided the highest level of accuracy. As a result, when compared to machine learning models majority voting obtained good accuracy. The comparison of various machine learning models used for detecting exoplanets was depicted in Figure 3. The accuracies of Decision Tree, SVC, Random Forest, KNN, and MLP were calculated, and the top four models with high accuracy (SVC, Random Forest, KNN, and MLP) were selected for majority voting. The highest accuracy was obtained through majority voting, which was 99.97%, whereas Decision Tree had the lowest accuracy of 93.4%. Random Forest, KNN, MLP, and SVC achieved accuracies of 99.88%, 98.47%, 98.65%, and 97.81%, respectively. To put it simply, the combination of the best performing models resulted in the highest accuracy of exoplanet detection.

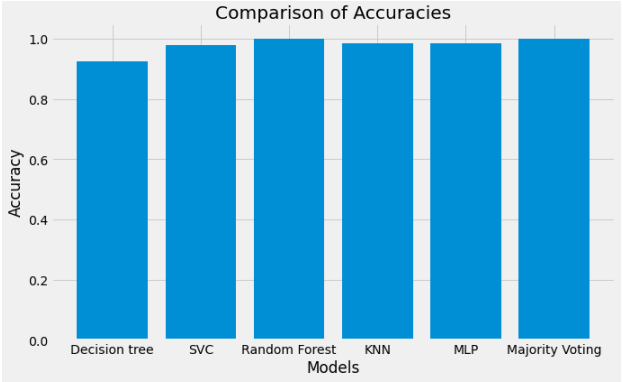

Fig. 3. Test Accuracies of Different Machine Models

By contrasting real and expected values, a confusion matrix evaluates the precision of a categorization model. Figure 4 shows the Confusion Matrix for a Random Forest, whereas Figure 5 illustrates the Confusion Matrix for a K-Nearest Neighbours (KNN) model. Figure 6 demonstrates the Confusion Matrix for a decision tree model, Figure 7 depicts the Confusion Matrix for a SVM model, and Figure 8 displays the Confusion Matrix for a Multilayer Perceptron (MLP) model. Figure 9 showcases the Confusion Matrix for a Majority Voting model.

Table I. Accuracy of different Models

Model	Accuracy	F1_score	Precision	Recall
Random Forest	99.88	99.88	99.88	99.88
KNN	98.47	98.47	98.52	98.47
Decision Tree	93.40	93.39	93.94	93.40
SVC	97.81	97.81	97.90	97.81
MLP	98.65	98.65	98.69	98.65
Majority Voting	**99.97**	**99.97**	**99.97**	**99.97**

The Precision-Recall Curve (PRC) and Receiver Operating Characteristic Curve (ROC) are used to illustrate the trade-off between true and false positive rates. The PRC examines the model's precision and recall at various thresholds, whereas the ROC depicts the link between sensitivity and specificity. Here, Figures 10 and 11 depict the ROC curve and the PRC curve of majority voting, respectively.

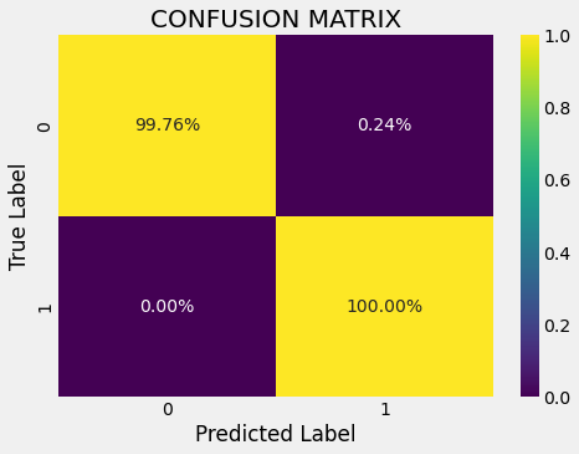

Fig. 4. Random Forest confusion matrix.

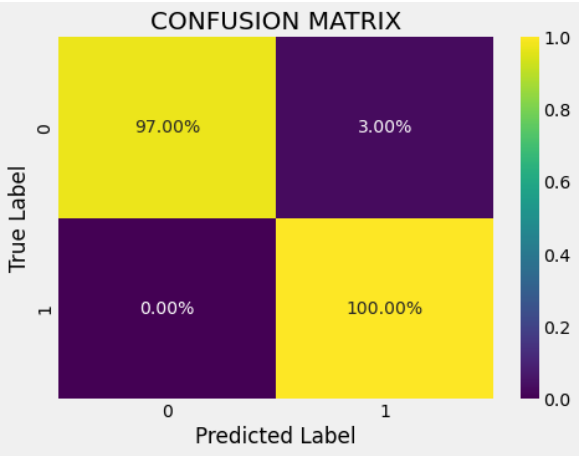

Fig. 5. KNN confusion matrix

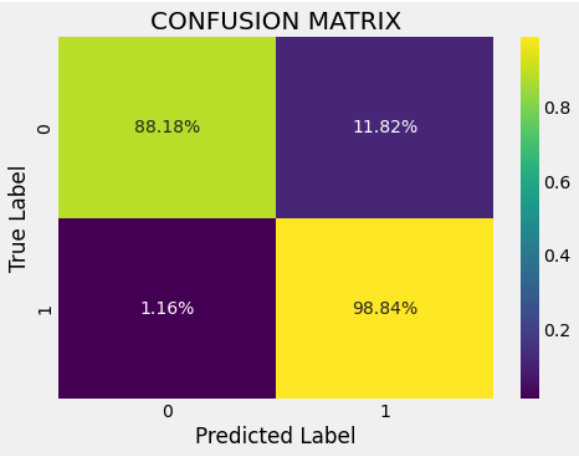

Fig. 6. Decision Tree confusion matrix

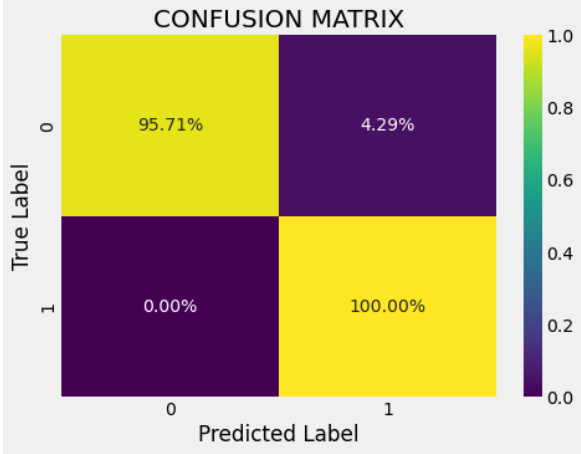

Fig. 7. SVC confusion matrix

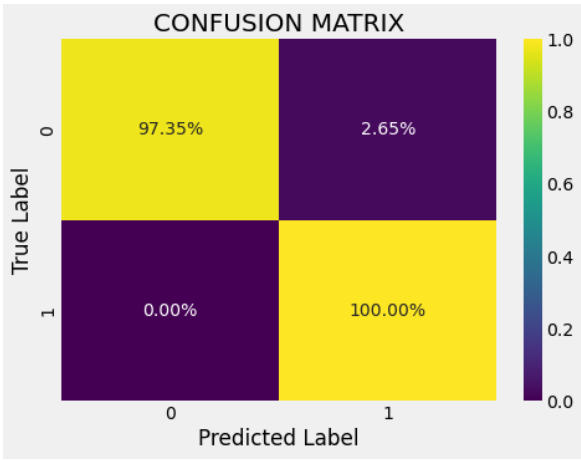

Fig. 8. MLP confusion matrix

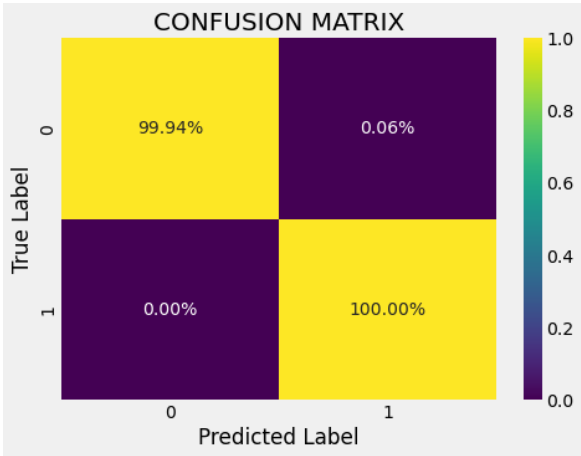

Fig. 9. Majority Voting confusion matrix

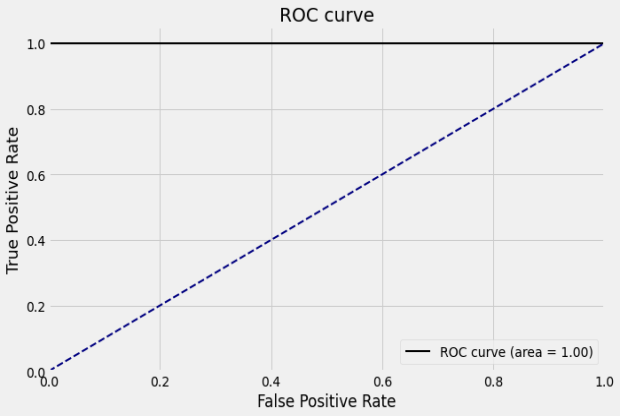

Fig. 10. Majority Voting ROC Curve

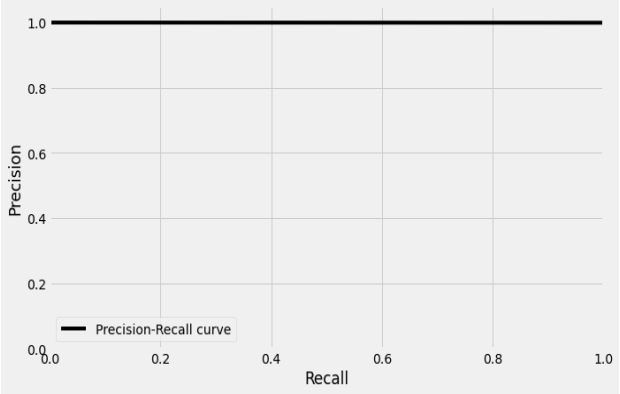

Fig. 11. Majority Voting PRC Curve

TABLE II. COMPARING WITH EXISTING MODELS

Author/Model	Accuracy (%)
George Clayton Sturrock et al [12]	98.96
J. Shallue & Andrew Vanderburg [13]	91.7
Liang Yu et. al [14]	97.4
Pattana Chintarungruangchai & Ing-Guey Jiang [15]	99
Majority Voting	**99.97**

Table II displays the exoplanet detection accuracy using machine learning and deep learning methods. We can see that compared to all other related works; we obtained the best precision of 99.97%. Since no one else has previously used this approach to discover exo-planets, we have used the majority voting method to determine accuracy.

V. CONCLUSION

We propose a machine learning approach for the automatic classification of conceivable possibilities for the Kepler transiting planet. Our method makes use of the majority voting model to accurately distinguish between transiting exoplanets and various false positives, including eclipsing binaries, instrument artifacts, and stellar variability with high accuracy. Our algorithm reliably prioritizes real planet candidates over false positives on our test dataset, resulting in a classification accuracy of 99.97%. The planet candidate identification process used by the Kepler mission and upcoming exoplanet surveys may be considerably more accurate and efficient thanks to this technique. In this study, five machine learning models, namely Random Forest (RF), Multilayer Perceptron (MLP), K-Nearest Neighbour (KNN), Decision Tree, SVM and were employed to determine the accuracy of each individual. The four models with the highest accuracy were then chosen using a majority voting process, which are RF, KNN, SVC, and MLP. Through the aforementioned process, an accuracy rate of 99.97% was obtained, which surpassed the performance of all other machine learning models used in previous studies. According to these results, the majority voting method utilising RF, KNN, SVC, and MLP may be able to improve the accuracy of individual identification tasks. Overall, the study's findings show that using multiple machine learning models and majority voting can help with individual recognition tasks by increasing accuracy levels. The present study employed various techniques, including exploratory data analysis, feature engineering, SMOTE, and ensemble learning, to create a highly accurate machine learning model. The study emphasized the significance of preparing the data and using appropriate methods to handle imbalances in the data. The ensemble learning technique was found to be effective in enhancing the model's performance.

In the future, researchers could explore alternative feature engineering and ensemble learning techniques to further improve model accuracy. Moreover, research efforts could concentrate on developing new strategies to address imbalanced data, which could lead to even more robust models.

REFERENCES

[1] Bugueno, M., Mena, F., & Araya, M. (2018). Refining exoplanet detection using supervised learning and feature engineering. In *2018 XLIV Latin American Computer Conference (CLEI)* (pp. 278-287). IEEE.

[2] Jara-Maldonado, M., Alarcon-Aquino, V., Rosas-Romero, R., Starostenko, O., & Ramirez-Cortes, J. M. (2020). Transiting exoplanet discovery using machine learning techniques: a survey. *Earth Science Informatics*, *13*, 573-600.

[3] Schanche, N., Cameron, A. C., Hébrard, G., Nielsen, L., Triaud, A. H. M. J., Almenara, J. M., ... & Wheatley, P. J. (2019). Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys. *Monthly Notices of the Royal Astronomical Society*, *483*(4), 5534-5547.

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

[5] Priyadarshini, I., & Puri, V. (2021). A convolutional neural network (CNN) based ensemble model for exoplanet detection. *Earth Science Informatics*, *14*, 735-747.

[6] Jin, Y., Yang, L., & Chiang, C. E. (2022). Identifying exoplanets with machine learning methods: a preliminary study. *arXiv preprint arXiv:2204.00721*.

[7] Dong, G., & Liu, H. (Eds.). (2018). *Feature engineering for machine learning and data analytics*. CRC Press.

[8] Matchev, K. T., Matcheva, K., & Roman, A. (2022). Unsupervised Machine Learning for Exploratory Data Analysis of Exoplanet Transmission Spectra. *The Planetary Science Journal*, *3*(9), 205.

[9] Ofman, L., Averbuch, A., Shliselberg, A., Benaun, I., Segev, D., & Rissman, A. (2022). Automated identification of transiting exoplanet candidates in NASA Transiting Exoplanets Survey Satellite (TESS) data with machine learning methods. *New Astronomy*, *91*, 101693.

[10] Aydoğan, Koray. "Exoplanet Detection by Machine Learning with Data Augmentation." arXiv preprint arXiv:2211.15577 (2022).

[11] Malik, Abhishek, Benjamin P. Moster, and Christian Obermeier. "Exoplanet detection using machine learning." Monthly Notices of the Royal Astronomical Society 513.4 (2022): pp- 5505-5516.

[12] Sturrock, G.C., Manry, B. and Rafiqi, S., 2019. Machine learning pipeline for exoplanet classification. SMU Data Science Review, 2(1), p.9.

[13] 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), p.94.

[14] Yu, L., Vanderburg, A., Huang, C., Shallue, C.J., Crossfield, I.J., Gaudi, B.S., Daylan, T., Dattilo, A., Armstrong, D.J., Ricker, G.R. and Vanderspek, R.K., 2019. Identifying exoplanets with deep learning. III. Automated triage and vetting of TESS candidates. The Astronomical Journal, 158(1), p.25.

[15] Chintarungruangchai, P. and Jiang, G., 2019. Detecting exoplanet transits through machine learning techniques with convolutional neural networks. Publications of the Astronomical Society of the Pacific, 131(1000), p.064502