



A convolutional neural network (CNN) based ensemble model for exoplanet detection

Ishaani Priyadarshini¹ · Vikram Puri²

Received: 28 August 2020 / Accepted: 25 January 2021 / Published online: 15 February 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract

Exoplanet detection is an extremely active research topic in astronomy. Researchers in the past have attempted to detect exoplanets using conventional methods like Radial Velocity, Transit Method, Gravitational Microlensing, Direct Imaging, Polarimetry, Astrometry, etc. While the approaches undertaken for all these studies vary, many of the research works conducted are based on the change in flux (light intensity). Based on the same characteristic, we explore yet another method of detecting exoplanets in space, using Artificial Intelligence. We rely on several machine learning algorithms like Decision Trees, Support Vector Machines, Logistic Regression, Random Forest Classifier, Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), as baseline algorithms and introduce an Ensemble-CNN model to draw out comparisons between the different machine learning models. The performance of the models has been evaluated using parameters like Accuracy, Precision, Sensitivity, and Specificity. Our results denote that the proposed Ensemble-CNN model performs relatively better for detecting exoplanets with an accuracy of 99.62%. The research will be useful in the fields of Astronomy as well as Artificial Intelligence and would be of substantial importance to physicists, cosmologists, scientists, researchers, academicians, industry experts, and machine learning experts who work in areas related to (or closely related to) exoplanet detection.

Keywords Exoplanet detection · Flux (light intensity) · Artificial intelligence · Machine learning · Convolutional neural networks (CNN) · Ensemble

Introduction

The world of astronomy is expansive and profound. One of the biggest curiosities of the human mind is investigating what lies outside earth and finding celestial bodies in space. For many years, astronomers and researchers observed the night sky and movement of planets and stars to identify specific stars, galaxies, and planets. While stars and galaxies are popular studies among physicists and astronomers, the topic of exoplanet detection gathers more curiosity among researchers

and the general public as there is a great curiosity to find extraterrestrial life or habitable conditions for supporting life. The research topic also explores the information with respect to the atmosphere and composition of such planets. It gives researchers an idea about how solar systems are formed, and the habitable zones they embrace. As we know, planets orbit the sun and have enough mass to converge into a sphere. Moreover, their gravitational force is significant in their orbits. An exoplanet is a planet that is not a part of our Solar System. Their masses are restricted to 13 Jupiter masses (Khan et al. 2017), and they are usually not dwarf planets (Zapatero Osorio et al. 2000). Some of the exoplanets that have been discovered over the last few decades are Kepler-186f, Kepler-16b, CoRoT 7b, and Kepler 22b (Quintana 2014; Doyle 2019; Barnes et al. 2010; Neubauer et al. 2012). The exoplanetary study has relied on the Kepler Space Telescope, Hubble Space Telescope, CoRoT satellite, Transiting Exoplanet Survey Satellite (TESS), NASA Spitzer Space Telescope, etc. for decades (Jara-Maldonado et al. 2020). Scientists and Researchers in the field of astronomy have tried to detect exoplanets using conventional methods like:

✉ Vikram Puri
purivikram@duytan.edu.vn

Ishaani Priyadarshini
ishaani@udel.edu

¹ Department of Electrical and Computer Engineering, University of Delaware, Newark, USA

² Center of Visualization and Simulation, Duy Tan University, Da Nang, Vietnam

Radial velocity, a method that analyzes the Doppler shift effect in the host star, due to the mutual gravity between an exoplanet and host star (Cornachione et al. 2019).

Transit Method, an exoplanet passing between the observer and its host star, produces a transit, which can be explored using light curves (Zaleski et al. 2019).

Gravitational Microlensing, massive objects distort the change in direction of light which leads to a gravitational lensing effect on the light of the star (Treu et al. 2012).

Direct imaging, a technique concerned with spatially resolving the exoplanet and its host star, such that images can be obtained from exoplanets (Kane et al. 2019).

Polarimetry, light reflected off the atmosphere of a planet leads to light waves interacting with atmospheric molecules, the atmosphere becomes polarized (Ren et al. 2019).

Astrometry, a method that measures the position of a star and observes the change in position in time (Lacour et al. 2019).

Since the exoplanets are not a part of the solar system, they are believed to be orbiting another star, which casts its light on them. Based on the intensity of the light (flux), various studies have been conducted for exoplanet detection.

Technology is growing at a rampant rate, and a lot of this may be attributed to the acceleration in Artificial Intelligence, which finds its use in many fields like computer systems (Pritam et al. 2019; Jha et al. 2019a, b, c), the energy industry (Puri et al. 2019), cybersecurity (Priyadarshini and Cotton 2020; Tuan et al. 2019; Priyadarshini 2018; Priyadarshini et al. 2019; Priyadarshini and Cotton 2019), healthcare (Priyadarshini et al. 2020; Dansana et al. 2020), finance (Lu 2019), and others (Jha et al. 2019a, b, c; Patro et al. 2020). As we know, most of the research works on Exoplanet Detection tremendously rely on the intensity of light. However, the challenge is the high contrast between the host star and its companions. Stellar variations, noise due to outliers, discontinuities within the light curves, weak signals, false alarm rates, and the plethora of information obtained are some other challenges in detecting exoplanets (Jara-Maldonado et al. 2020; Mullally et al. 2016). Since machine learning can handle large amounts of data and make decisions based on trends and patterns, it would be interesting to consider machine learning algorithms for exoplanet detection based on the intensity of light (flux). In this paper, we take into account eight artificial intelligence techniques including machine learning algorithms for the purpose of exoplanet detection. The techniques used are Decision Trees, Support Vector Machines, Logistic Regression, Convolutional Neural Networks (CNN), Random Forest Classifier, and Multilayer Perceptron (MLP). Moreover, the Ensemble-CNN based approach has also been proposed for the study, which makes it the first and only one of its kind based on the state of art. The overall performance of the models has been evaluated using performance measures like Accuracy, Precision, Sensitivity,

and Specificity. While there are many research papers that perform studies on exoplanet detection, this is the first of its kind to address the issue extensively using Artificial Intelligence techniques. The research will be beneficial to physicists, cosmologists, scientists, researchers, academicians, industry experts, and machine learning experts who work in areas related to (or closely related to) exoplanet detection.

The rest of the paper is organized as follows. Section 2 depicts Materials and Methods in which we discuss prior work pertaining to the research and the artificial intelligence techniques that we have considered for the study. Section 3 manifests the datasets taken into consideration and the experimental analysis. In section 4, we present the results including performance evaluation and comparative analysis. Finally, in Section 5 we discuss the Conclusion and the Future Work.

Materials and methods

This section has been divided into two parts. Section 2.1 describes the prior work related to exoplanet detection, and Section 2.2 discusses the Artificial Intelligence techniques taken into account for the study.

Related works

(Melchior et al. 2018) suggested a methodology for exoplanet detection using WFIRST (Wide-Field Infrared Survey Telescope) Diffraction Spikes using a point spread function (PSF) model. The study asserts that when stars were observed using WFIRST Wide Field Camera, they saturated the detector and produced strong diffraction spikes. The analytical model constructed propagates Poisson noise from the star and thermal thermal emission from the telescope. Using the bluest filters acquires the best precision due to diffraction spikes being the narrowest.

(Flasseur et al. 2018) proposed another approach for exoplanet detection by Direct Imaging, The Unsupervised Patch Based algorithm (PAch COvariance or PACO) is used for modeling spatial correlation and background non-stationarity. A time-series observation manifests the distribution of background patches, for improving the robustness and fluctuations. The background model is refined using sub-pixel location and flux of the exoplanet. Detected sources are removed from data and control of false alarms serves as the stopping criteria. Detection maps and ROC curves have been used for performance evaluation.

(Pearson et al. 2018) performed a study using Artificial intelligence for the detection of exoplanets. A deep learning methodology has been proposed for the same. The CNN proposed can detect exoplanets in noisy time series data with better accuracy in comparison to the least-squares method.

Deep nets can be generalized and allow data to be evaluated from different time series. Although it is based on interpolation, it does not compromise the performance.

(Schanche et al. 2019) suggested four machine learning approaches for exoplanet transit detection. The machine learning algorithms taken into consideration are Linear SVC (Support Vector Classifier), SVC, Logistic Regression, K-Nearest Neighbor, Random Forest, and Convolutional Neural Networks (CNN). It was observed that the accuracy for CNN is the best with approximately 90%.

(Chintarungruangchai and Jiang 2019) suggested the use of Convolutional Neural networks for exoplanet detection. Five deep learning models including one Multilayer Perceptron and four CNN have been used for the study. The evaluation has been performed based on Training, Validation, and Testing Processes, Signal Noise ratios, Transit phase positions, and Folding periods. The study asserts that models with folding achieve greater accuracy.

(Dattilo et al. 2019) recommended identifying exoplanets using deep learning. The study also claims to identify two new super earths using neural networks. The characteristics considered for the study are Light Curve Production, Transit Search, Labeling Threshold Crossing Events, etc. The evaluation has been conducted taking into consideration accuracy, precision, recall, false positives, and area under the curve. The accuracy of the model is approximately 98%.

(Sturrock et al. 2019) proposed yet another approach for exoplanet classification using machine learning pipeline. While the study takes into account several classification models and datasets to derive the probability that an observation is an exoplanet, Random Forest Classifier has been observed to be the most optimal classifier with an accuracy of 98%. The study also asserts that 968 observations have a greater than 95% probability of being an exoplanet.

(Yu et al. 2019) suggested a deep learning methodology for identifying exoplanets. An existing neural network has been modified to Kepler candidates, while the first neural network is applied to TESS (Transiting Exoplanet Survey Satellite) data. Or triage mode, the average accuracy is 97.4%, while in vetting mode the accuracy is 97.7%. Likewise, the average precision in triage mode is 97%, and for vetting mode is 69.3%.

(Mathur et al. 2020) presented a Deep Learning approach for identifying exoplanets and the likelihood of their habitability. The study is based on automatically classifying transit signals into exoplanets or non-exoplanets with respect to different methods of signal processing. A CNN model has been trained for the classification. The habitability of the planets has been studied using the Naive Bayes algorithm for planetary characteristics.

(Singh et al. 2020) propounded the use of CNN (deep learning) for predicting viable exoplanets. The study investigates the best neural network based on the number of layers,

filter size, and data. This is followed by a prediction to find if a planet is statistically qualified to be an exoplanet or if it is a false positive.

Based on the literature survey, we observed that the problem has been addressed using conventional astronomical methods as well as artificial intelligence techniques in the past. Based on the Related Works, we also observe that

1. Exoplanet Detection continues to be a burning problem for decades. We conduct a study on the topic in this article.
2. There are a handful of papers that address the issue using Artificial Intelligence (AI), so Exoplanet detection using AI is largely unexplored. We explore it yet again using some AI methods.
3. Most of the articles perform the study using baselines algorithms. There is no article that proposes an AI method explicitly for exoplanet detection. We propose an **Ensemble-CNN** model for the same.

Methods

Exoplanet Detection has always been a topic of interest among astronomers, researchers, scientists, and physicists. While a variety of conventional methods have been used in the past for the same, the advancement of artificial intelligence over the last few years has made this a machine learning problem, which is now not confined to only physics and planetary studies. The last few years have witnessed several machine learning techniques being applied to the study for defining the probability of detecting an exoplanet. In this paper, we rely on the intensity of light (flux) for the same. When a planet passes between Earth and its host star, it leads to a dip in the brightness or intensity of the observed star, which is likewise referred to as Flux. As we know, planets themselves do not emit light. However, if a star is observed for several months, one can observe regular dimming of the flux. This is indicative of the fact that there is an orbiting body around the star, which may potentially belong to its system, and maybe an exoplanet. Using Artificial Intelligence Techniques, including Machine Learning algorithms, it may be possible to classify whether the orbiting body is an exoplanet or not. The following are some of the Artificial Intelligence methods considered for this research:

Decision trees

Decision Trees refer to tree-like graphs with nodes depicting the point from which we choose an attribute and ask a question (Kingsford and Salzberg 2008). While edges are known to represent the answers to the question, leaves represent the actual output or class label. Decision trees find their use in nonlinear decision making with simple linear decision

surfaces (Elavarasan et al. 2018). Decision trees are more or less concerned with creating training models that may be useful in predicting the class or value of the target variable. This supervised learning algorithm relies on simple decision rules acquired from training data. In order to predict a class label for a record, we consider beginning at the root of the tree. The values of the root attribute and the record attribute are compared. Based on the comparison, the branch corresponding to the value is followed, and the next node is considered. The decision tree consists of a root node, which is split into decision nodes. A decision node may further split into decision nodes or terminal nodes (leaf node). Each node in the tree can be thought of as a test case for some attribute. Similarly, each edge descending from the node can be thought of as possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new node (see Fig. 1).

Support vector machines

Support vector machines are machine learning algorithms based on supervised learning. They present a useful pathway for analyzing data with respect to classification and regression analysis. During the 1990s, SVM was developed with the target of providing non-linear solutions for the analysis of data. SVM is one of the most widely used machine learning algorithms (Ben-Hur et al. 2001; Chang et al. 2018). This is because, with a minimum number of features, the ability of SVM is more as compared to other model techniques. The SVM model is relatively robust compared to the error of the model. Moreover, the computational time of SVM is lower than other models like neural networks. Finally, the efficiency of SVM is better than most models. SVM architecture can be

depicted in three different layers: The Input Layer, the Hidden Layer, and the Output Layer. Layer 1 or input layer incorporates training instances that are connected to the hidden layer for processing the prediction of learning (see Fig. 2). This layer is further connected to the output layer.

Logistic regression

Logistic regression is a machine learning algorithm that is used for conducting regression analysis, given the dependent variable is dichotomous (Menard 2002). It is a predictive analysis, which is used to describe data and determine the relationship between dependent binary variables and independent variables. Thus, we can say that the dependent variable is a binary variable that is identified by data 1 (for yes, happy, success, etc.) or 0 (for no, sad, failure, etc.). The logistic regression model predicts $P(Y = 1)$ as a function of X . The advantage of logistic regression over linear regression is that it can directly predict probabilities. Further, it preserves the marginal probabilities of training data. There are some assumptions associated with logistic regression. It does not consider variables that do not have meaning and require large sample sizes. It is mostly used to predict output which is binary.

Random Forest classifier

The Random Forest classifier is basically an ensemble tree-based learning algorithm (Ho 1995). Thus, it consists of a large number of individual decision trees that operate as an ensemble. The classifier takes into account a set of decision trees from a randomly selected subset of the training set such that votes from different decision trees are aggregated to determine the final class of the test object (Chakriswaran et al.

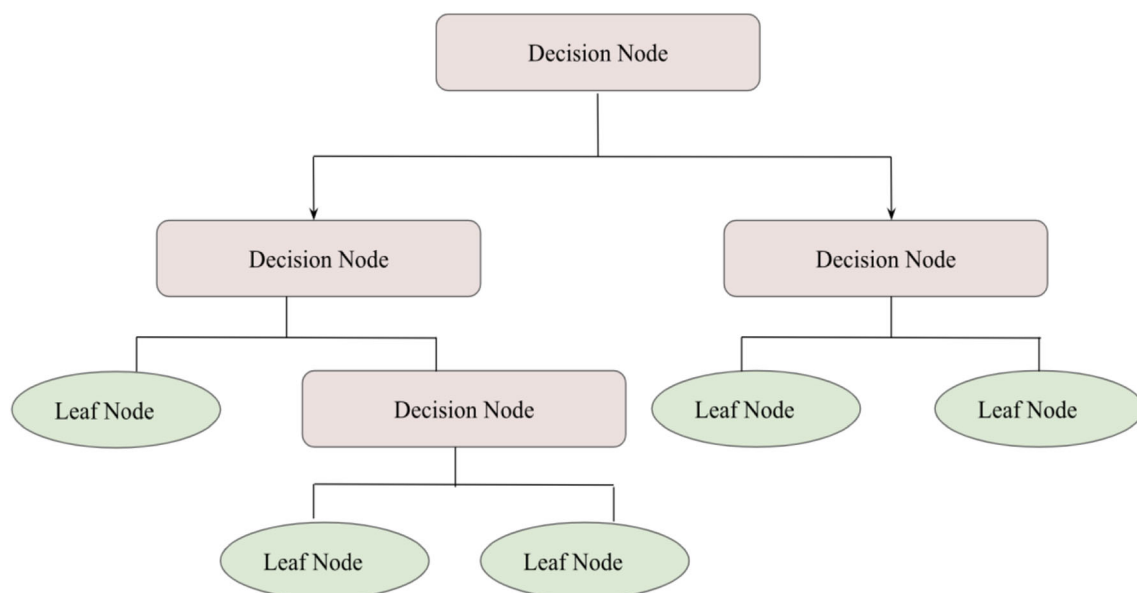


Fig. 1 Decision Tree

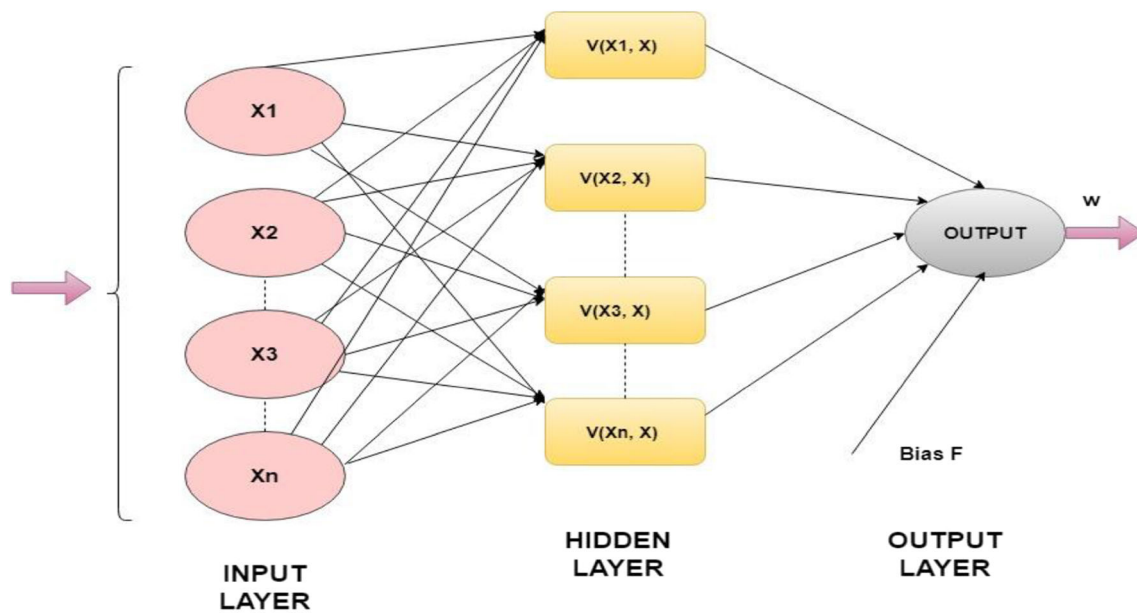


Fig. 2 SVM Architecture

2019). Since every tree has a class prediction, the class with the highest or maximum number of votes becomes the model's prediction. This is also indicative of the fact that if there are a dominant number of relatively uncorrelated models that operate as a committee, then they will outperform the individual constituent models (see Fig. 3). The classifier works using the following steps

- Selecting random samples from a given dataset.
- Constructing a decision tree for every sample
- Obtain the prediction result for each decision tree.

- Perform voting for every predicted result.
- Choose the most voted prediction result to be the final prediction result.

Multilayer perceptron

In the Perceptron Algorithm, a single perceptron is multiplied by a weight and a bias is added. However, in Multilayer Perceptron (MLP), there are multiple linear layers. A three-

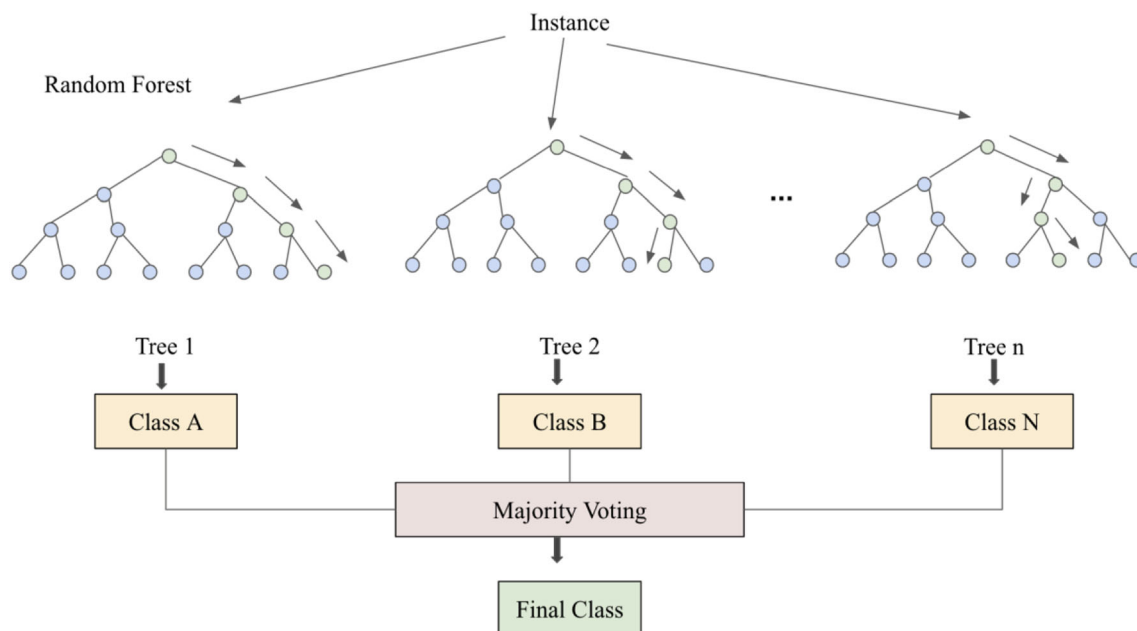


Fig. 3 Random Forest Classifier

layer network incorporates an input layer, a hidden layer, and an output layer. Data is fed to the input layer and output is obtained from the output layer. The number of hidden layers can be modified depending on the problem (Tang et al. 2015). MLP is fully connected which means that every perceptron is connected to every other perceptron (see Fig. 4). Fully connected MLP includes too many parameters which can lead to weights becoming unmanageable (Quek et al. 2019). This can result in inefficiency and redundancy, and overfitting. Compared to MLP, Convolutional Neural Networks (CNN) are more robust and can handle more parameters.

Convolutional neural networks (CNN)

A Convolutional Neural Network belongs to a class of deep Neural networks and uses filters for performing convolution operations. The convolutional layer is the basic building block of a CNN (Simard et al. 2003; Srinivasan et al. 2018). The layer's parameters comprise a lot of learnable filters, which consolidate a little responsive field, however reach out through the full profundity of the input volume. During the forward pass, every filter is convolved over the width and height of the input volume, processing the dot product between the entries of the filter and the input and delivering a 2-dimensional activation map. Thus, the network learns filters that actuate when it identifies some particular kind of feature at some spatial position in the information. Convolutional layers convolve the information and its outcome to the following layer. This is exactly like the reaction of a neuron in the visual cortex to a

particular stimulus. Each convolutional neural procedure information for its specific receptive field. A CNN is made out of a few sorts of layers. The Convolutional layer is responsible for creating a feature map for anticipating the class probabilities for each feature. This is done by applying a filter that scans the entire image. A pooling layer scales down the measure of data the convolutional layer created for each feature and keeps up the most significant information. The fully connected input layer is concerned with smoothing the outputs produced by past layers to transform them into a single vector that can be utilized as input for the following layer. The fully connected layer is responsible for applying weights over the input generated by the feature analysis. The purpose of this function is to predict an accurate label. A fully connected output layer is responsible for generating the last probabilities for deciding a class for the image.

Proposed solution

In this study, the proposed system architecture has five levels (see Fig. 5): 1) Exoplanet Dataset 2) Pre-Processing of Data 3) Data Categorization 4) Application of Machine Learning Models 5) Performance Evaluation. The architecture and the steps have been detailed as follows.

1. **Exoplanet Dataset:** The first layer of the proposed system is responsible for collecting data from the exoplanet dataset based on the identification. The collected data is in the form of a .csv file and consists of 3198 features.

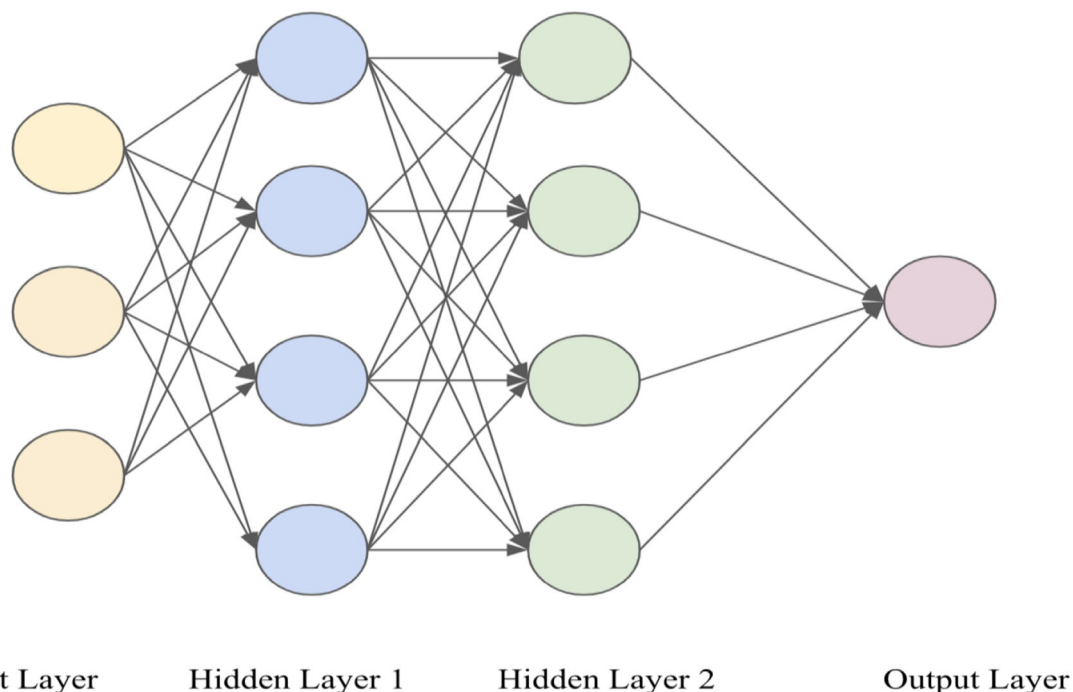


Fig. 4 Multilayer Perceptron

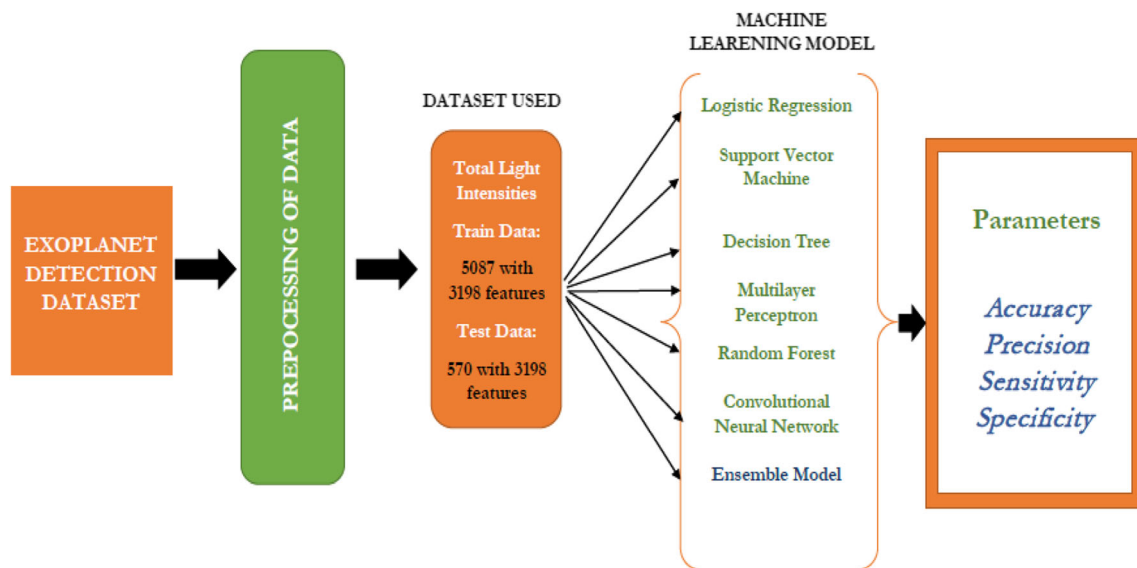


Fig. 5 Architecture of the Proposed Solution

2. **Data Pre-Processing:** In this step, data filtered and processed to remove all NaN values. After removal of unwanted data, the data is split as per requirement
3. **Data Categorization:** Filtered data is split into train and test categories. We obtain 5067 samples with 3198 features and 570 samples with 3198 features respectively. Now the dataset is prepared for the application of machine learning models.
4. **Machine Learning models:** Six pre-existing machine learning models such as logistic regression, support vector machine, decision tree, multilayer perceptron, random forest, and CNN models are applied. This is followed by the application of a proposed Ensemble-CNN model which is also integrated with the system.
5. **Evaluation:** Evaluation is one of the major for assessing the performance of the models. The dataset has been already split into training and testing data for evaluating the model performance. In this study, four different parameters are used such as accuracy, precision, sensitivity, and specificity.

Ensemble CNN model

Stacking is a way of combining different machine learning models. Stacking (Wang et al. 2011) introduces the concept of meta learner that alternates the concept of Voting, which is widely used by the bagging. The main role of the stacking is to identify or classify the reliable model and meta learner and it helps to discover a method of integrating the results of the best base-learner. Level-1 models and Level-0 models are the predictions of the metamodel input and base model respectively. Stacked learners are basically used for classification such that an occurrence is provided into a Level-0 model and every

occurrence predicts a class value. These predictions append to level-1 and combine to construct the final prediction. Training the stacking model is quite a tricky concept but not as difficult as it sounds and the training steps of the stacking model are similar to that of k-fold validation. In this Ensemble technique, the dataset is categorized into two sets: Train and Test set. However, the test set is not used in the process of training. The training set is further divided into k-number of folds. These folds contain N/k number of points if the input dataset contains N points of data. With the use of the M number of models, it predicts the value of fold, and the M -Number of predictions is obtained from the N/k points of data. These predictions can be used as input for the meta learner and final results can be predicted from the metal learner. Figure 6 represents the proposed ensemble model. In this model, support vector machines, random forest, multilayer perception, and decision trees are the base learners and CNN is the meta learner.

Experimental analysis

In this section, we will discuss the datasets used for the study followed by the implementation.

Datasets

In order to detect an exoplanet, we have considered changes in flux or light intensity of the star. The dataset has been taken from Dataset. The data incorporates the change in flux of several thousand stars such that each star is associated with a binary label of either 1 or 2. Label 1 denotes a non-exoplanet star while Label 2 denotes an exoplanet star. A total of 3197 light intensities have been considered in the dataset ranging from FLUX1 to FLUX3197. The dataset exists as a .csv file,

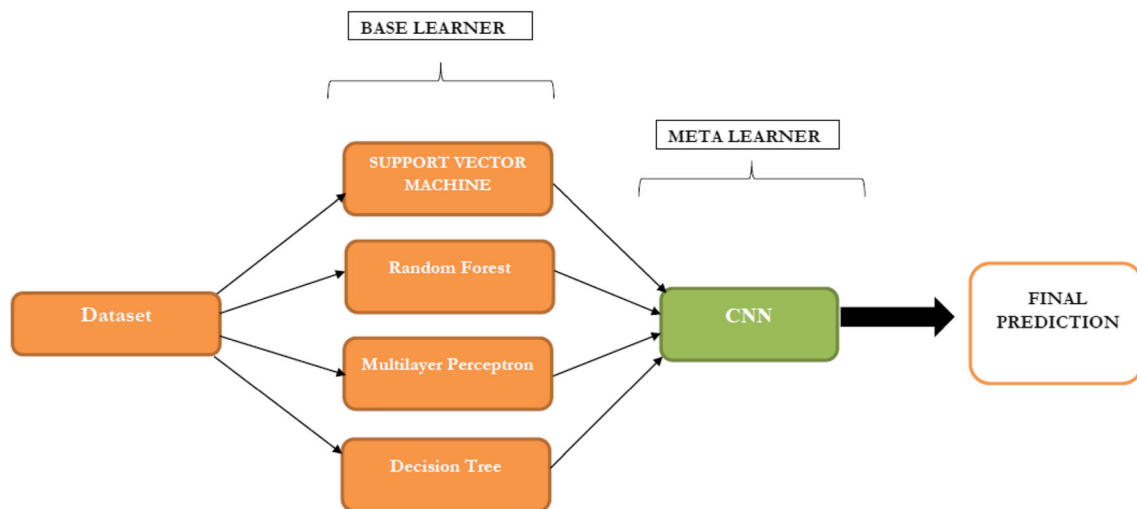


Fig. 6 Proposed Ensemble Model

and the experimental analysis using the Artificial Intelligence techniques has been performed using python. The train set has 5087 rows or observations, and 3198 columns or features. While Column 1 is the label vector, Columns 2–3198 depict flux values over time. In this set, there are 37 confirmed exoplanet-stars and 5050 non-exoplanet-stars. For the test set, there are 570 rows or observations and 3198 columns or features. Again, Column 1 is the label vector, whereas, for Columns 2, 3198 are the flux values over time. There are 5 confirmed exoplanet-stars and 565 non-exoplanet-stars in this set.

Implementation

In order to conduct the experimental analysis, we have considered the following steps.

In Fig. 7, we observe that there are 3197 (values) + 1 column, therefore in total, there are 3198 columns.

a. Treating Missing or Duplicate values

Following the analysis for treating or missing values, we find that there are no missing or duplicate data. Hence, prediction models might be accurate.

- b. Treating outliers. We observe that the outlier shown is very high, but the number of outliers is only 1, hence we exclude that data point.
- c. Observing the relationship between variables

1 Scatter Matrix

Scatter plots are used for data visualization and depict the relationship between two variables. We observe that for the first 7 light intensities, the graph is almost linear (see Fig. 8).

2 Pair plots

Pair plots can be used for visualizing the relationship between variables. The matrices of relationships seen here for the first 7 light intensities are almost linear (see Fig. 9).

	LABEL	FLUX.1	FLUX.2	FLUX.3	FLUX.4	FLUX.5	FLUX.6	FLUX.7	FLUX.8	FLUX.9	...
count	5087.000000	5.087000e+03	5.087000e+03	5.087000e+03	5.087000e+03	5.087000e+03	5.087000e+03	5.087000e+03	5.087000e+03	5.087000e+03	...
mean	1.007273	1.445054e+02	1.285778e+02	1.471348e+02	1.561512e+02	1.561477e+02	1.469646e+02	1.168380e+02	1.144983e+02	1.228639e+02	...
std	0.084982	2.150669e+04	2.179717e+04	2.191309e+04	2.223366e+04	2.308448e+04	2.410567e+04	2.414109e+04	2.290691e+04	2.102681e+04	...
min	1.000000	-2.278563e+05	-3.154408e+05	-2.840018e+05	-2.340069e+05	-4.231956e+05	-5.975521e+05	-6.724046e+05	-5.790136e+05	-3.973882e+05	...
25%	1.000000	-4.234000e+01	-3.952000e+01	-3.850500e+01	-3.505000e+01	-3.195500e+01	-3.338000e+01	-2.813000e+01	-2.784000e+01	-2.683500e+01	...
50%	1.000000	-7.100000e-01	-8.900000e-01	-7.400000e-01	-4.000000e-01	-6.100000e-01	-1.030000e+00	-8.700000e-01	-6.600000e-01	-5.600000e-01	...
75%	1.000000	4.825500e+01	4.428500e+01	4.232500e+01	3.976500e+01	3.975000e+01	3.514000e+01	3.406000e+01	3.170000e+01	3.045500e+01	...
max	2.000000	1.439240e+06	1.453319e+06	1.468429e+06	1.495750e+06	1.510937e+06	1.508152e+06	1.465743e+06	1.416827e+06	1.342888e+06	...
8 rows x 3198 columns											

Fig. 7 Data Description. a. Description of Data

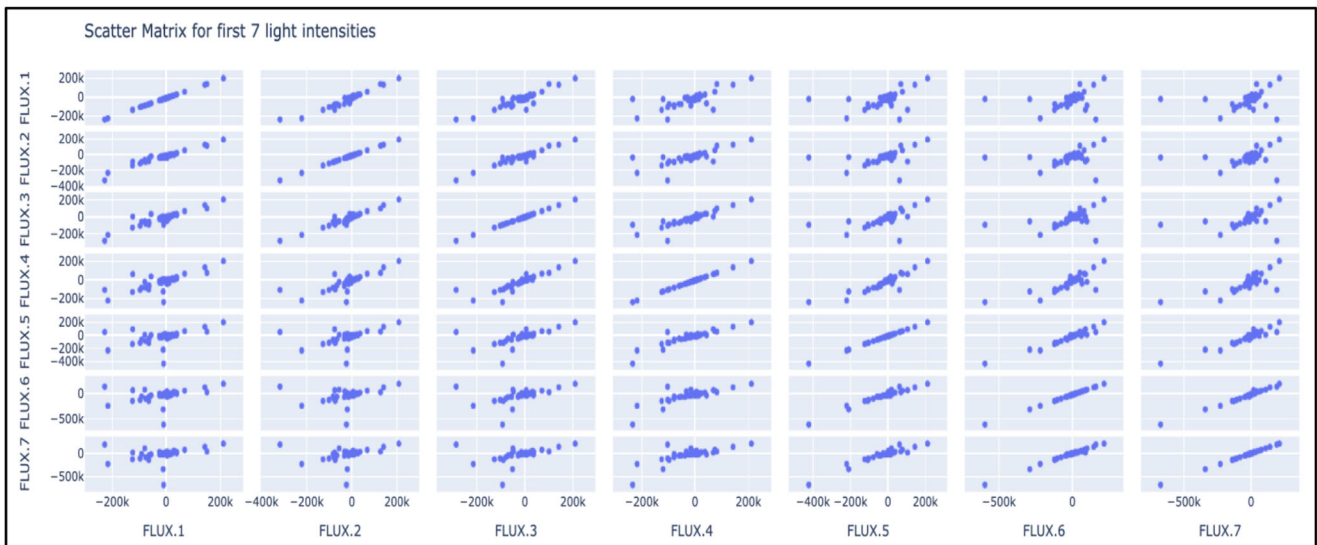


Fig. 8 Scatter Matrix for first 7 Light Intensities

a. Following this, the Artificial Intelligence Techniques and Machine Learning Algorithms were used for exploratory analysis.

- Logistic Regression
- Support Vector Machine
- Decision Tree
- Multilayer Perceptron
- Random Forest
- Convolutional Neural Networks
- Ensemble-CNN Model

Results

This section is divided into two parts. In the first part, we present the performance evaluation of the Artificial Intelligence techniques considered for exoplanet detection. The second section highlights the comparative analysis of our work with previous works, followed by discussion,

Performance evaluation

Performance Evaluation is one of the significant aspects of the machine learning process. For detecting whether an observation is an exoplanet or not, we have deployed several machine learning algorithms and artificial intelligence techniques. There is a need to determine model performance by means of evaluation. Performance Evaluation has been conducted using four parameters namely,

a. **Accuracy:** Accuracy is a metric that may be used for evaluating classification models. It emphasizes how often an

algorithm classifies data correctly. It may be defined as the number of correctly predicted data points with respect to the total number of data points. Given a confusion matrix, accuracy may be defined as the sum of True Negative and True Positives combined over the sum of True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP).

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

b. **Precision:** Precision is a metric that is popularly used in classification, information retrieval, and pattern recognition. It may be defined as the number of relevant observations with respect to retrieved observations. Given a confusion matrix, Precision may be calculated by True Positives with respect to the total number of True Positives and False Negatives combined.

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

c. **Sensitivity:** Sensitivity metric is defined as the ratio of actual positive events that got predicted as positive. It is sometimes also referred to as recall. Given a confusion matrix, sensitivity may be calculated by True positive value with respect to the sum of True Positive and False Negative combined)

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

d. **Specificity:** Specificity metric is defined as the ratio of actual negatives that got predicted as negative. Given a

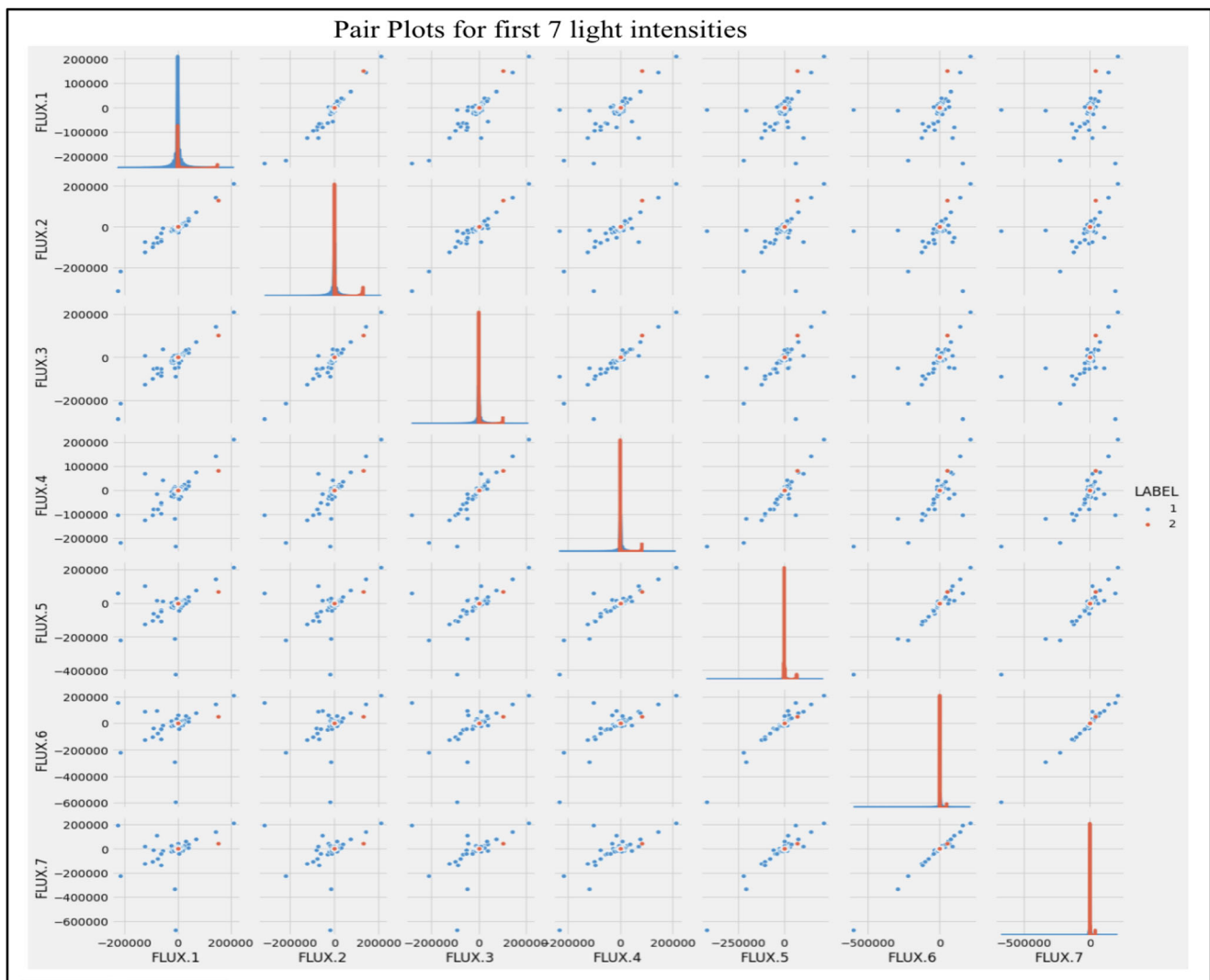


Fig. 9 Pair Plots for first 7 Light Intensities

confusion matrix, specificity would be calculated by True Negatives with respect to the sum of True Negatives and False Positives combined.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

As per Table 1, we evaluate the seven artificial models including the proposed model in this study. The accuracy of the logistic regression, support vector machine, decision tree, multilayer perceptron, random forest, CNN, and Ensemble-CNN models are 0.7842, 0.9862, 0.9456, 0.9812, 0.9713, 0.9132, and 0.9962 respectively. The precision values are

Table 1 Performance Evaluation of Methods Applied

Artificial Intelligence Methods	Accuracy	Precision	Sensitivity	Specificity
Logistic Regression	0.7842	0.9338	0.62	0.512
Support Vector Machine	0.9862	0.9737	0.9899	0.897
Decision Tree	0.9456	0.9828	0.90	0.982
Multilayer Perceptron	0.9812	1.00	0.80	1.00
Random Forest	0.9713	0.9846	0.968	0.935
Convolutional Neural Networks (CNN)	0.9132	0.9034	1.00	0.912
Proposed Ensemble-CNN model	0.9962	0.9867	0.9989	0.9872

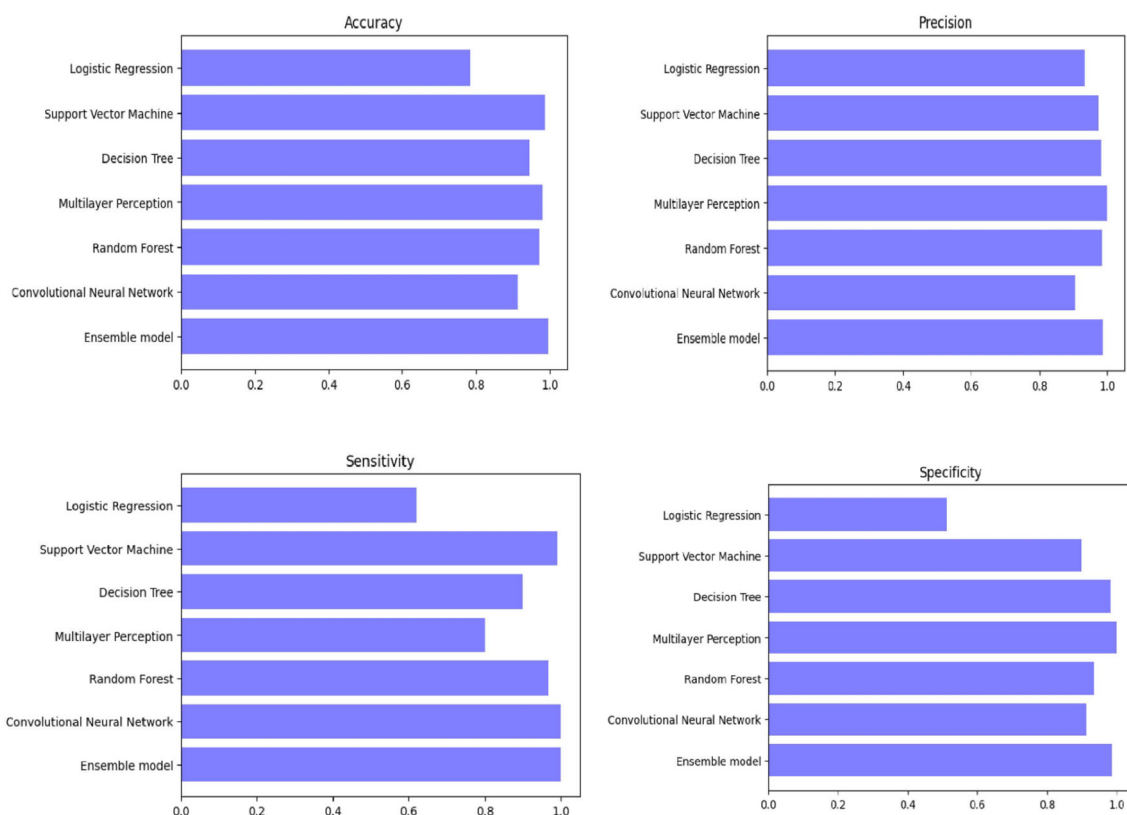


Fig. 10 Comparison of performance evaluation of the models with Ensemble Model

0.9338, 0.9737, 0.9828, 1.00, 0.9846, 0.9034 and 0.9867 respectively. Sensitivity and specificity are two parameters that assist in evaluating the trained model. The Value for the sensitivity and specificity are 0.62:0.512, 0.9899:0.897, 0.90:0.982, 0.80:1.00, 0.968:0.935, 1.00:0.912 and 0.9989:0.9872 respectively (see Fig. 10). The performance of the logistic regression is the least as compared to the other models. The Proposed Ensemble-CNN model performs

relatively better with respect to the other baseline models used in the study.

Comparative analysis

In this section, we present a comparative analysis of our work with existing works. The same is depicted in Table 2.

Table 2 Comparative analysis of Proposed ensemble model with other studies

Year and Author	Research	Methodology/ Parameters	Results
Mislis et al. 2018	Transiting Exoplanet Light Curves	Machine Learning Data Rejection Algorithm	Detection Efficiency ~80%
Zucker and Giryes 2018	Detecting periodic transits of exoplanets	Deep Learning	Sensitivity=0.94
Amin et al. 2018	Detecting Exoplanet Systems	Adaptive Neuro-Fuzzy System	Accuracy~81%
Zingales and Waldmann 2018	Retrieving Exoplanetary Atmosphere	Deep Convolutional Generative Adversarial Networks	Speed improvement over traditional retrievals by 300 times
Ansdehl et al. 2018	Improvised Exoplanet Transit Classification	Deep Learning	Increase in model accuracy and average precision by 2.0%–2.5%
Chintarungruangchai and Jiang 2019	Detecting Exoplanet Transits	Machine Learning and CNN	Accuracy ~98%
Jara-Maldonado et al	Survey on Transiting Exoplanet Discovery	Machine Learning	Highest Accuracy obtained by Random Forests which is 97.82
Our Proposed Work, 2020	Exoplanet Detection using AI techniques	Machine learning models and proposed Ensemble CNN model	Accuracy=99.62%

We observe that our model outperforms the previously used models for exoplanet detection with an accuracy of 99.62%.

Conclusion and future work

Exoplanet detection is a fascinating area of research and has been studied using various methods. This research work explores yet another technique of detecting exoplanets using robust artificial intelligence models. Some conventional methods used for the study are the Transit Method, Radial velocity, Direct Imaging, Gravitational Microlensing, etc. In this paper, we used the intensity of light (flux) and Artificial Intelligence techniques including Machine Learning algorithms like Logistic Regression, Support Vector Machines, Decision Tree, Multilayer Perceptron, Random Forest, and Convolutional Neural Networks to draw a comparison between different models for exoplanet detection. Moreover, we also proposed an Ensemble-CNN model for the same. While most of them depict satisfactory results, our proposed model outperforms them all with an accuracy of 99.62%.

In the future, we would like to analyze more machine learning models and artificial intelligence techniques for detecting exoplanets. Moreover, the study has some limitations as it is not easy to detect light reflected from a planet's atmosphere. Therefore, it would be interesting to explore other methods and combine them with Artificial Intelligence techniques for exoplanet detection.

Author's contribution Conception and Design of Work: Ishaani Priyadarshini and Vikram Puri, Data Collection: Ishaani Priyadarshini, Data Analysis and Interpretation: Ishaani Priyadarshini and Vikram Puri, Drafting the Article: Ishaani Priyadarshini, Critical Revision of the Article: Vikram Puri, Final Approval of the Version to be submitted: Vikram Puri.

Funding This research received no external funding.

Declaration

We declare that this manuscript is original, has not been published before, and is not currently being considered for publication elsewhere.

Ethics approval NA

Consent to participate NA

Conflict of interest The authors declare no conflict of interest.

References

Amin RA, Khan AT, Raisa ZT, Chisty N, Samiha Khan S, Khaja MS, Rahman RM (2018) Detection of exoplanet systems in Kepler light curves using adaptive Neuro-fuzzy system. In 2018 international conference on intelligent systems (IS) (pp. 66–72). IEEE

Ansdell M, Ioannou Y, Osborn HP, Sasdelli M, Smith JC, Caldwell D et al (2018) Scientific domain knowledge improves exoplanet transit classification with deep learning. *Astrophys J Lett* 869(1):L7

Barnes R, Raymond SN, Greenberg R, Jackson B, Kaib NA (2010) CoRoT-7b: super-earth or super-Io? *Astrophys J Lett* 709(2):L95–L98

Ben-Hur A, Horn D, Siegelmann HT, Vapnik V (2001) Support vector clustering. *J Mach Learn Res* 2(Dec):125–137

Chakriswaran P, Vincent DR, Srinivasan K, Sharma V, Chang CY, Reina DG (2019) Emotion AI-driven sentiment analysis: a survey, future research directions, and open issues. *Appl Sci* 9(24):5462

Chang CY, Srinivasan K, Chen SJ, Chang MS, Sharma V (2018) An efficient SVM based lymph node classification approach using intelligent communication ant Colony optimization. *J Med Imaging Health Informatics* 8(5):1077–1086

Chintarunguangchai P, Jiang G (2019) Detecting exoplanet transits through machine-learning techniques with convolutional neural networks. *Publ Astron Soc Pac* 131(1000):064502

Cornachione MA et al (2019) A full implementation of Spectro-perfectionism for precise radial velocity exoplanet detection: a test case with the MINERVA reduction pipeline. *Publ Astron Soc Pac* 131(1006):124503

Dansana D, Kumar R, Adhikari JD, Mohapatra M, Sharma R, Priyadarshini I, Le DN (2020) Global forecasting confirmed and fatal cases of COVID-19 outbreak using autoregressive integrated moving average model. *Frontiers in public health*, 8

Dataset. Kaggle, Kepler Labelled Time Series Data. <https://www.kaggle.com/keplersmachines/kepler-labelled-time-series-data>

Dattilo A, Vanderburg A, Shallue CJ, Mayo AW, Berlind P, Bieryla A et al (2019) Identifying Exoplanets with Deep Learning. II. Two New Super-Earths Uncovered by a Neural Network in K2 Data. *Astronom J* 157(5):169

Doyle LR (2019) The discovery of “Tatooine”: Kepler-16b. *New Astron Rev* 84:101515

Elavarasan D, Vincent DR, Sharma V, Zomaya AY, Srinivasan K (2018) Forecasting yield by integrating agrarian factors and machine learning models: a survey. *Comput Electron Agric* 155:257–282

Flasseur O, Denis L, Thiébaud E, Langlois M (2018) An unsupervised patch-based approach for exoplanet detection by direct imaging. In 2018 25th IEEE international conference on image processing (ICIP) (pp. 2735–2739). IEEE

Ho TK (1995). Random decision forests. In proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278–282). IEEE

Jara-Maldonado M, Alarcon-Aquino V, Rosas-Romero R, Starostenko O, Ramirez-Cortes JM (2020) Transiting exoplanet discovery using machine learning techniques: a survey. *Earth Sci Inform* 13:573–600. <https://doi.org/10.1007/s12145-020-00464-7>

Jha S, Kumar R, Chiclana F, Puri V, Priyadarshini I (2019a) Neutrosophic approach for enhancing quality of signals. *Multimed Tools Appl*:1–32

Jha S, Kumar R, Priyadarshini I, Smarandache F, Long HV (2019b) Neutrosophic image segmentation with dice coefficients. *Measurement* 134:762–772

Jha S, Kumar R, Abdel-Basset M, Priyadarshini I, Sharma R, Long HV (2019c) Deep learning approach for software maintainability metrics prediction. *Ieee Access* 7:61840–61855

Kane SR, Dalba PA, Li Z, Horch EP, Hirsch LA, Horner J, Wittenmyer RA, Howell SB, Everett ME, Butler RP, Tinney CG, Carter BD, Wright DJ, Jones HRA, Bailey J, O'Toole SJ (2019) Detection of planetary and stellar companions to neighboring stars via a combination of radial velocity and direct imaging techniques. *Astron J* 157(6):252

Khan MS, Stewart Jenkins J, Yoma N (2017) Discover- ing new worlds: a review of signal processing methods for detecting exoplanets from

- astronomical radial velocity data. *IEEE Signal Process Mag* 34: 104–115. <https://doi.org/10.1109/MSP.2016.2617293>
- Kingsford C, Salzberg SL (2008) What are decision trees? *Nat Biotechnol* 26(9):1011–1013
- Lacour S et al (2019) First direct detection of an exoplanet by optical interferometry-astrometry and K-band spectroscopy of HR 8799 e. *Astronomy Astrophys* 623:L11
- Lu Y (2019) Artificial intelligence: a survey on evolution, models, applications and future trends. *J Manag Anal* 6(1):1–29
- Mathur, S., Sizon, S., & Goel, N. (2020) Identifying exoplanets using deep learning and predicting their likelihood of habitability. In *advances in machine learning and computational intelligence* (pp. 369–379). Springer, Singapore
- Melchior P, Spergel D, Lanz A (2018) In the crosshair: Astrometric exoplanet detection with WFIRST's diffraction spikes. *Astron J* 155(2):102
- Menard S (2002) *Applied logistic regression analysis* (Vol. 106). Sage
- Mislić D, Pyrzak S, Alsubai KA (2018) TSARDI: a machine learning data rejection algorithm for transiting exoplanet light curves. *Mon Not R Astron Soc* 481(2):1624–1630
- Mullally F, Coughlin JL, Thompson SE, Christiansen J, Burke C, Clarke BD, Haas MR (2016) Identifying false alarms in the Kepler planet candidate catalog. *Publ Astron Soc Pac* 128(965):074502
- Neubauer D, Vrtala A, Leitner JJ, Firneis MG, Hitzemberger R (2012) The life supporting zone of Kepler-22b and the Kepler planetary candidates: KOI268. 01, KOI701. 03, KOI854. 01 and KOI1026. 01. *Planet Space Sci* 73(1):397–406
- Patro SGK, Mishra BK, Panda SK, Kumar R, Long HV, Taniar D, Priyadarshini I (2020) A hybrid action-related K-nearest neighbour (HAR-KNN) approach for recommendation systems. *IEEE Access* 8:90978–90991
- Pearson KA, Palafox L, Griffith CA (2018) Searching for exoplanets using artificial intelligence. *Mon Not R Astron Soc* 474(1):478–491
- Pritam N, Khari M, Kumar R, Jha S, Priyadarshini I, Abdel-Basset M, Long HV (2019) Assessment of code smell for predicting class change proneness using machine learning. *IEEE Access* 7:37414–37425
- Priyadarshini I (2018). Features and architecture of the modern cyber range: a qualitative analysis and survey (Doctoral dissertation, University of Delaware)
- Priyadarshini I, Cotton C (2019, October) Internet memes: a novel approach to distinguish humans and bots for authentication. In *proceedings of the future technologies conference* (pp. 204–222). Springer, Cham
- Priyadarshini I, Cotton C (2020) Intelligence in cyberspace: the road to cyber singularity. *J Exp Theoretic Artificial Intell* 1–35
- Priyadarshini I, Wang H, Cotton C (2019, October) Some Cyberpsychology techniques to distinguish humans and bots for authentication. In *proceedings of the future technologies conference* (pp. 306–323). Springer, Cham
- Priyadarshini I, Mohanty P, Kumar R, Son LH, Chau HTM, Nhu VH, Ngo P, Tien Bui D (2020) Analysis of outbreak and global impacts of the COVID-19. In *healthcare* (Vol. 8, no. 2, p. 148). Multidisciplinary digital publishing institute
- Puri V, Jha S, Kumar R, Priyadarshini I, Abdel-Basset M, Elhoseny M, Long HV (2019) A hybrid artificial intelligence and internet of things model for generation of renewable resource of energy. *IEEE Access* 7:111181–111191
- Quek SG, Selvachandran G, Munir M, Mahmood T, Ullah K, Son LH et al (2019) Multi-attribute multi-perception decision-making based on generalized T-spherical fuzzy weighted aggregation operators on neutrosophic sets. *Mathematics* 7(9):780
- Quintana E (2014). Kepler 186f—the first earth-sized planet orbiting in habitable zone of another star
- Ren D, Ranganathan M, Christian DJ (2019) A host-star calibration based Polarimeter for earth-like exoplanet imaging. *Publ Astron Soc Pac* 131(1005):115004
- Schanche N, Cameron AC, Hébrard G, Nielsen L, Triaud AHMJ, Almenara JM, Alsubai KA, Anderson DR, Armstrong DJ, Barros SCC, Bouchy F, Boumis P, Brown DJA, Faedi F, Hay K, Hebb L, Kiefer F, Mancini L, Maxted PFL, Pallé E, Pollacco DL, Queloz D, Smalley B, Udry S, West R, Wheatley PJ (2019) Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys. *Mon Not R Astron Soc* 483(4):5534–5547
- Simard PY, Steinkraus D, Platt JC (2003) Best practices for convolutional neural networks applied to visual document analysis. In *Icdar* (Vol. 3, no. 2003)
- Singh G, Gawane S, Prasad A, Wagaskar K (2020) Modeling CNN for best parameter investigation to predict viable exoplanets. In *advanced computing technologies and applications* (pp. 591–607). Springer, Singapore
- Srinivasan K, Sharma V, Jayakody DNK, Vincent DR (2018, December) D-ConvNet: deep learning model for enhancement of brain MR images. In *basic & Clinical Pharmacology & Toxicology* (Vol. 124, pp. 3–4). 111 RIVER ST, HOBOKEN 07030-5774. WILEY, NJ
- Sturrock GC; Manry B; Rafiqi, Sohail (2019) Machine Learning Pipeline for Exoplanet Classification," *SMU Data Science Review*: Vol. 2 : No. 1 , Article 9
- Tang J, Deng C, Huang GB (2015) Extreme learning machine for multi-layer perceptron. *IEEE Trans Neural Networks Learn Syst* 27(4): 809–821
- Treu T, Marshall PJ, Clowe D (2012) Resource letter GL-1: gravitational lensing. *Amer J Phys* 80:753–763. <https://doi.org/10.1119/1.4726204> arXiv:1206.0791
- Tuan TA, Long HV, Kumar R, Priyadarshini I, Son NTK (2019) Performance evaluation of botnet DDoS attack detection using machine learning. *Evol Intel*:1–12
- Wang G, Hao J, Ma J, Jiang H (2011) A comparative assessment of ensemble learning for credit scoring. *Expert Syst Appl* 38(1):223–230
- Yu L, Vanderburg A, Huang C, Shallue CJ, Crossfield IJ, Gaudi BS et al (2019) Identifying Exoplanets with Deep Learning. III. Automated Triage and Vetting of TESS Candidates. *Astronom J* 158(1):25
- Zaleski SM, Valio A, Marsden SC, Carter BD (2019) Differential rotation of Kepler-71 via transit photometry mapping of faculae and starspots. *Mon Not R Astron Soc* 484(1):618–630
- Zapatero Osorio MR et al (2000) Discovery of young, isolated planetary mass objects in the σ orionis star cluster. *Science*
- Zingales T, Waldmann IP (2018) Exogan: retrieving exoplanetary atmospheres using deep convolutional generative adversarial networks. *Astron J* 156(6):268
- Zucker S, Gyries R (2018) Shallow transits—deep learning. I. Feasibility study of deep learning to detect periodic transits of exoplanets. *Astronom J* 155(4):147