Detecting Exoplanets Using CNN Algorithm

Pranav Chavan

M.Sc. in Computing
in Big Data Analytics
and Artificial
Intelligence

2022-2023

Ollscoil Teicneolaíochta an Atlantaigh Dún na nGall

Atlantic Technological University Donegal

Department of Computing, ATU Donegal, Port Road, Letterkenny, Co. Donegal, Ireland.

Author: Pranav Chavan

Supervised by: John Conaghan

A thesis submitted in partial fulfilment of the requirements for the

Master of Science in Computing in Big Data Analytics and Artificial Intelligence

Submitted to Atlantic Technological University
Arna chur isteach chuig Ollscoil Teicneolaiochta an Atlantaigh

September 2022-2023

Declaration

I hereby certify that the material, which l now submit for assessment on the programmes of study leading to the award of Master of Science in Computing in Big Data Analytics and Artificial Intelligence, is entirely my own work and has not been taken from the work of others except to the extent that such work has been cited and acknowledged within the text of my own work. No portion of the work contained in this thesis has been submitted in support of an application for another degree or qualification to this or any other institution. I understand that it is my responsibility to ensure that I have adhered to ATU's rules and regulations.

I hereby certify that the material on which I have relied on for the purpose of my assessment is not deemed as personal data under the GDPR Regulations. Personal data is any data from living people that can be identified. Any personal data used for the purpose of my assessment has been pseudonymised and the data set and identifiers are not held by ATU. Alternatively, personal data has been anonymised in line with the Data Protection Commissioners Guidelines on Anonymisation.

I consent that my work will be held for the purposes of education assistance to future students and will be shared on the ATU Donegal (Computing) website (www.lyitcomputing.com) and Research THEA website (https://research.thea.ie/). I understand that documents once uploaded onto the website can be viewed throughout the world and not just in Ireland. Consent can be withdrawn for the publishing of material online by emailing Jade Lyons; Head of Department at Jade.Lyons@atu.ie to remove items from the ATU Donegal Computing website and by email emailing Denise McCaul; Systems Librarian at denise.mccaul@atu.ie to remove items from the Research THEA website. Material will continue to appear in printed formats once published and as websites are public medium, ATU cannot guarantee that the material has not been saved or downloaded.

31/08/2023

Signature of Candidate

Date

Acknowledgements

I would like to express my sincere appreciation to Mr. John Conaghan, my dedicated supervisor for my dissertation. His consistent support, invaluable guidance, and unwavering encouragement have been pivotal throughout this research endeavor. The timely and constructive feedback he provided played a crucial role in enhancing the overall quality and depth of the study.

I extend my gratitude to the esteemed faculty at ATU Letterkenny whose unwavering assistance and wealth of knowledge have been indispensable. Their commitment to offering continual support, sharing essential insights, and providing access to cutting-edge technologies and resources has been instrumental in successfully concluding this dissertation.

I am also deeply thankful to the participants of this study. Their willingness to contribute their insights and experiences has greatly enriched the research.

Lastly, my heartfelt thanks go to my family and friends for their unending belief in me. Their constant inspiration and encouragement have been a driving force behind this accomplishment. This dissertation stands as a testament to their unwavering support and motivation.

.

Abstract

The search for exoplanets remains a top priority in the field of astronomy, driven by the ambitious aim of understanding the deep complexity of our enormous cosmos and finding potentially habitable worlds. Traditional analytical approaches have experienced serious restrictions in light of amazing advances in telescopic technologies.

To accomplish this feat, two types of data were used: planet transit image data and the extremely extensive Kepler dataset, which contains tabular information about numerous exoplanetary properties. Convolutional Neural Networks (CNN), VGG16, VGG19, Support Vector Machine (SVM), Random Forest, and a Stacking classifier were among the machine learning models tested.

The empirical outcomes of this research effort revealed that models based on the Kepler dataset outperformed those trained on picture data. When applied to the Kepler dataset, the CNN model demonstrated an accuracy rate of 82.60%. The picture data categorization models did not perform well, producing a consistent accuracy of only 57.14%.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Background of the study

The use of deep learning to the detection of exoplanets using data from NASA's Kepler spacecraft is extremely important in the area of astronomy, particularly in the context of exoplanet discovery. With a growing reliance on automated ways to sift through massive volumes of astronomical data, the necessity for efficient and accurate detection procedures has become critical. Exoplanets, celestial bodies that orbit stars outside our solar system, have attracted scientists' interest for decades. The NASA-launched Kepler spacecraft has played a critical role in this endeavour by providing an enormous dataset on probable exoplanetary transits. This information has proven to be a vital resource for academics attempting to unravel the secrets of these faraway worlds (Ansdell et al., 2018).

Deep learning, an area of machine learning, has emerged as an exciting tool in a variety of scientific fields, including astronomy. Recent advances have demonstrated its ability to automate the classification of Kepler transit signals, efficiently differentiating between actual exoplanets and false positives with exceptional accuracy (Shallue & Vanderburg, as quoted by Ansdell et al., 2018). Incorporating scientific domain knowledge into deep learning models has also been proven to improve their performance, particularly in recognizing weak signals in noisy data. This combination of experience and cutting-edge technology has created new opportunities for exoplanet detection and characterization (Ansdell et al., 2018).

While classic methods of exoplanet detection have proven to be efficient, they frequently necessitate significant manual labor and may not scale well with the increasing amount of data. Deep learning, on the other hand, provides a scalable answer to this problem. Convolutional neural networks (CNNs) and other models have showed potential in detecting planetary transits in Kepler Telescope light curves, presenting a more efficient and automated approach to exoplanet discovery (Cuellar et al., 2021). The integration of both actual and synthetic data in the training process has been investigated to increase the performance of these models even more. This combination has proved its ability to improve transit identification in real-world light curves, demonstrating the power of deep learning in exploiting different datasets for more accurate findings (Cuellar et. al., 2021).

Given the ever-changing landscape of astronomical research and the enormous potential of automated approaches, the goal of this work is to dive deeper into the possibilities of deep learning in exoplanet finding utilizing rich data from the Kepler mission. This research not only contributes to the larger scientific debate by improving our understanding of these faraway worlds, but it also prepares the way for the advancement of intelligent systems in space research. The combination of human skill and artificial intelligence holds enormous promise for solving the universe's riddles.

1.2 Research Motivation

The expanse of the universe, with its billions of stars and probably even more planets, confronts astronomers with an awe-inspiring challenge. Understanding the vastness and complexity of

the universe is a challenging task. The Kepler project, a ground-breaking effort in exoplanet exploration, has transformed our understanding of distant planetary systems. Kepler has discovered and cataloged thousands of exoplanets, helping to solve the mystery of celestial bodies outside our solar system.

This enormous accomplishment, however, comes with a considerable cost. The data produced by the Kepler mission is nothing short of incredible. The sheer volume of information is mind-boggling, and manually processing this data is not only time-consuming but also prone to human mistake. It is a difficult task that demands painstaking attention to detail as well as a great deal of patience.

To overcome this issue, there is an urgent need for automated technologies that can interpret and evaluate the massive amounts of data produced by Kepler. Deep learning is a powerful technology that has demonstrated its worth in a variety of disciplines. Scientists can potentially uncover hidden patterns buried inside massive amounts of data by utilizing the powers of deep learning.

The beauty of deep learning is its capacity to identify subtle links and correlations that standard analysis approaches may have missed. It offers the potential to uncover exoplanets that were previously overlooked or disregarded. These undiscovered treasures contain significant insights into the origin and evolution of planetary systems.

Deep learning can give information on the circumstances required for the existence of life beyond our own planet by locating these elusive exoplanets. It opens up a world of possibilities in which we can investigate the diversity of planetary systems and discover the secrets of the universe. Deep learning breakthroughs not only broaden our knowledge but also feed the human imagination, igniting a sense of wonder and curiosity.

Finally, the immensity of the universe and the wealth of data provided by the Kepler mission pose an enormous challenge to astronomers. Manually analyzing this data is both time-consuming and error-prone. Deep learning, with its extraordinary powers, provides a possible solution to this problem. Scientists can use deep learning to find hidden patterns, leading to the discovery of more exoplanets and providing crucial insights into the circumstances required for life. It is an intriguing frontier with the potential to transform our understanding of the universe and our role within it.

1.3 Research Question

This work attempts to answer a question by exploiting the vast dataset offered by the Kepler mission and drawing on the great potential of deep learning.

"How does deep learning, specifically convolutional neural networks, compare to standard methods in discovering exoplanets from Kepler's transit data?"

Furthermore, *can the incorporation of actual and synthetic data during the training phase improve the model's performance in identifying genuine exoplanetary transits in the presence of noisy data?*

1.4 Research Outline

This chapter introduced readers to the basic terminologies, the motivation for conducting this study and the aim of this study. The rest of the thesis is structured as follows:

The literature review chapter deals with the review of past studies done in the field of exoplanet detection in order to identify the gap that can be addressed through this research. Following which in depth methodology for conducting the study has been presented which includes the data collection, pre-processing and modeling techniques adopted for the study. The results of the implementation of the system are discussed in detail in chapter 4 following which the conclusion of the study is provided along with the future avenues of research.

2 Literature Review

In the realm of exoplanet detection, a comprehensive understanding of existing research is indispensable for shaping the trajectory of advancements. The literature review serves as the cornerstone upon which new insights are built, providing a panoramic view of the current state of knowledge, identifying gaps, and offering a foundation for further exploration. This chapter embarks on a journey by critically analyzing and synthesizing a diverse range of perspectives, methodologies, and findings, this literature review aims to elucidate the key themes, trends, and controversies that form the backdrop against which the current research unfolds.

2.1 Convultional Neural Network

A “convultional neural network” (CNN) can be highlighted as a type of artificial neural network which are significantly used within the aspect of image recognition as well as processing. CNN also can be highlighted as a significantly powerful tool through which the images can be recognised, however, it significantly requires millions of labelled data points for training. CNN as well as deep learning played a significantly essential role in the “Exoplanet detection mission” of “NASA Kepler” (Jais *et al*., 2019). It also can be said that the deep learning process can help to determine the images by monitoring the brightness of stars. It also has been observed that with the implementation of CNN as well as a deep learning method, NASA can develop the probability within their detection mission which can ensure the improvement within their operations. However, it also can be said that the deep learning process can be implemented through which the organisation can enable the chances to determine the images of 100,000 stars simultaneously. With the utilization of various layers of interconnected modes, the deep learning method can enable the chances to extract the “hierarchical representations” from the raw input in the aspect of deep learning within the recognition of images (Bouwmans *et al.*, 2019). In short, the process of deep learning can provide significant opportunities through which various fields can get help.

InceptionNet is another effective deep learning model that has been designed for the classification of images as well as developed for the high-scale visual recognition challenge. However, it has been seen that it is known for the development of the usage of “inception modules, blocks of layers” which design to learn about the combination of local as well as global features within the context of the input information that can be beneficial for the organisations (Irudayaraj, 2022).

2.2 InceptionNet

In the case of “NASA”, “InceptionNet” can be mentioned as the primary utilisation within the computer vision tasks which can be applied within the aspect of the Kepler mission. The mentioned mission significantly focuses on the analysis of light curves in the context of stars through which the potential exoplanets can be determined by NASA. It also has been observed

that InceptionNet is significantly designed for image classification as well as in the analysis of the information within the aspect of Kepler through which all the inputs of the information which can be collected by NASA can be analysed with the help of this technique (Irudayaraj, 2022). InceptionNet can enable the chances to evaluate the information which can be beneficial for the research in the context of an “Exoplanet detection mission” (Zhang and Guo, 2021). It can be highlighted as the effective model through which the information can be collected as well as successfully analysed which can help NASA to evaluate the data that can provide better outcomes.

2.3 Recurrent Neural Network

The RNN model is another effective approach within deep learning which can help with the process of neutral recognition as well as language processing. Before attention models, RNNs were the go-to solution for handling sequential data. A deep feedforward model may demand specific parameters for each component of the sequence (Fu *et al*., 2020). On the other side, it can be mentioned as the effective model by which sequential information can be handled as well as also improvement can be made witjhin the processing of the information. It has been observed that, unlike the traditional neural networks, the RNN model possesses the memory element which can enable them to process the information as per the sequence which can ensure a better development of the information as well as also a better analysis process within the aspect of the information which can be collected in the context of “Exoplanet detection mission” (Fu *et al*., 2020). RNN can be mentioned as a significantly effective model which can include the “processing of neutral language, speech recognition as well as machine translation and also the analysis within the aspect of time series”. It also can be observed that the mentioned effective approach can enable the chances to make predictions as well as can enable the generation of the outputs by engaging their memory and also incorporate the contextualization of the adjustment within the predictions (Fu *et al*., 2020).

In recent years, there has been a significant surge of interest in deep learning among the academic community, particularly in regard to models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). These models have demonstrated their efficacy in effectively extracting complex patterns from unprocessed data. The ramifications of these abilities have significant importance for NASA's ambitious "Exoplanet detection mission" as discussed by Jiang et al. (2019). Through the utilization of deep learning techniques, NASA has the potential to augment its likelihood of success in the identification and knowledge of exoplanets.

Nevertheless, NASA has access to a range of tools beyond deep learning. The application of XGBoost offers a potentially advantageous approach for enhancing the analysis of unprocessed data inside NASA's operational processes. XGBoost, originally developed with a focus on improving the performance of weak predictive models such as decision trees, exhibits a distinctive capability to produce exceptionally powerful predictive models. This characteristic renders it a suitable choice for the analysis of NASA's extensive collection of unprocessed data.

2.4 Ensemble Models

The success of XGBoost can be attributed to its capability to optimize gradients and fix faults caused by previous trees through sequential training. XGBoost guarantees the production of high-quality predictive models by reducing the loss function (Luo et al., 2021). In addition, XGBoost improves the accuracy of forecasts for NASA by capturing intricate feature relationships.

Categorical Boosting (Catboost) classifier on the other hand, can be highlighted as an effective algorithm for boosting which has been designed for the tasks that can be conducted through the machine learning process and is significantly similar to XGBoost. This algorithm can have the ability to handle the efficient categorical features through which the process of the information can be made significantly faster by the organizations to collect the appropriate information and obtain better outcomes within the operational process. The categorical features within the context can be mentioned as significantly common which are similar to real-world datasets (Kamran, 2021). These features significantly required preprocessing before usage in the "gradient boosting models". On the other side, it also can be said that "CatBoost" can enable the chances to handle the categorical features through the usage of significant innovative approaches that can be beneficial for the improvement of the operations for NASA (Kamran, 2021). It also provides better improvement possibilities which can ensure better growth in the aspect of an innovative approach named "ordered boosting". This approach can use several techniques which include statistical information as well as gradient-based optimization which can enable the chances to handle effective categorical variables which can ensure better processing of the information which can provide better outcomes within the operations of NASA that can be beneficial for the improvement of their performance as well as for the optimization of the effective information.

It has been seen that, in order to gradually extract higher-level features from the raw input, deep learning is a type of machine learning technique that employs many layers. For instance, in image processing, lower layers may recognize boundaries, while higher layers may recognise things that are important to people, such as numbers, letters, and faces. On the other side, it also can be considered an effective approach that can ensure data processing as well as the development of information that can be beneficial for NASA and its project. Based on the deep learning theory it can be mentioned that the effective approach can be determined which can ensure better improvement that can lead to efficiency within the aspect of the machine learning process. It also helps to identify the effectiveness in the context of the utilisation of operational information which can be processed through various models of the deep learning process (Gupta *et al.*, 2022). In addition to this, it can be mentioned that the effectiveness of the deep learning 90del can have significant effects on the optimisation of the information as well as in the effective approaches which are regarding the innovative development in the operational process of NASA. Based on the deep earning theories the development within data processing can be determined by the organisation which can enable the improvement within the operational process of the formation collection in the aspect of the machine learning process. By optimising the selected information the mentioned effectiveness can be developed

and also successful conduction can be made within the context of the operational process in data processing for NASA.

2.5 Literature review

The study made by Ofman et al., 2022.focuses on “The NASA Transiting Exoplanets Survey Satellite (TESS)” dataset is used to test a unique artificial intelligence (AI) method that combines numerous algorithms created by ThetaRay, Inc. and machine learning (ML) methodologies. Before using TESS data, the ML/AI ThetaRay system is first applied to “Kepler exoplanetary” data and validated with proven exoplanets. By using unsupervised and semi-supervised machine learning approaches, existing and new features of the data are built and employed in AI/ML analysis. These features are based on numerous observable characteristics. For further analysis the algorithm produces about 50 targets after being applied to 10,803 light curves of threshold crossing events (TCEs) produced by the TESS mission and obtained from the Mikulski Archive for Space Telescopes. Through additional manual vetting, they discover three new exoplanetary candidates. This study shows, for the first time, how certain coupled multiple AI/ML-based approaches can successfully classify TCEs quickly and automatically using a huge astrophysical dataset.

For years, researchers have searched for and found thousands of transiting exoplanets using information from “NASA's Kepler Space Telescope”. Kepler spotted stars during the extended K2 mission in a variety of sky areas throughout the ecliptic plane, indicating that these stars were located in distinct galactic settings. Astronomers are interested in finding out how the populations of exoplanets differ in these various environments. To do this, though, calls for a method that can automatically and objectively distinguish between exoplanets in these areas and false-positive signals that resemble transiting planet signals. Three study which is made by Dattilo et al., 2019. offer a technique for categorising these exoplanet signals using deep learning, a class of machine learning algorithms. In order to be able to recognise exoplanets in several K2 campaigns that occur in a variety of galactic settings, they updated a neural network that was previously used to identify exoplanets in the Kepler field. In order to determine if a specific potential exoplanet signal is actually created by an exoplanet or a false positive, they train a convolutional neural network dubbed AstroNet-K2. AstroNet-K2 has a 98% accuracy on the test set for classifying exoplanets and false positives.

In this study Vishwarupe et al., 2022 focus on the one significant step in this direction is “NASA's Kepler Mission”, which uses telescopes to survey the “Milky Way” galaxy and search for thousands of Earth-sized and other smaller planets in or near the habitable zone in order to identify the thousands or even millions of stars in the galaxy that may host such planets in orbit. Any new planet beyond the solar system that revolves around a star is an exoplanet. Finding new exoplanets gives us the opportunity to fully comprehend the mechanisms involved in planet creation. arduous task to extract potential exoplanets from the mission data using conventional techniques. This can be accomplished by comparing several algorithms, which reveals the benefits and drawbacks of each method when used to analyse particular types of data.

Here, Yu et al., 2019 demonstrates a deep learning model that can evaluate and vet TESS applicants with the help of a class of machine learning algorithms. The model, which is the first neural network to be trained and tested on actual TESS data, is a modification of an existing neural network created to automatically categorise Kepler candidates. The model can differentiate transit-like signals from stellar variability and instrument noise in triage mode with an average accuracy and precision of 97.0% and 97.4%, respectively. With the aid of freshly added scientific domain knowledge, the model is trained to identify only planet candidates in vetting mode, and it reaches an overall accuracy and precision of 69.3% and 97.8%.

This research which is made by Singh and Misra., 2020. Get focuses on the use of several machine learning methods to forecast the potentially habitable dispositions of exoplanets using NASA's Kepler data. There will also be a comparison of the effectiveness of various algorithms. The outcomes will be used to determine whether algorithms are appropriate for making predictions regarding exoplanets. It's time to detect exoplanets using machine learning. A Better understanding of planet habitability, star bodies, and the variety of exoplanets in the galaxy will result from this. The model may be improved upon as new data from space telescopes come in, increasing accuracy. The proposed approach will be able to identify exoplanet candidates as habitable or non-habitable based on data produced by various surface and satellite observatories.

Exoplanets have been discovered in the galaxy using NASA's Kepler Space Telescope. In this research, utilising data from the Kepler space telescope and its extended mission K2, they discuss expanding on some previous work on exoplanet identification using residual networks. This study conducted by Kumari, A., 2023. For intends to investigate how deep learning algorithms can aid in classifying the existence of exoplanets when there is less data available in one scenario and a wider range of data available in another. they suggest a Siamese architecture in addition to the conventional CNN-based approach, which is particularly helpful in handling classification in a low-data scenario. The average accuracy of classification for the ResNet and CNN algorithms was 86% for two classes and 68% for three classes.

With the help of “NASA's Kepler spacecraft”, the search for planets that could support life has advanced significantly. Around 4000 planets have been successfully found by the mission, but the manual review of this data is laborious and time-consuming, necessitating the development of more effective techniques for finding exoplanets in order to eliminate errors and false positives. The purpose of this project which is made by Sharma et al., 2023. is to categorise stars as exoplanets using data gathered by the Kepler satellite using machine learning methods. To do this, they intend to employ preprocessing techniques and appropriate classification algorithms to create an exact and ideal classifier, hence enhancing the process's competency.

The wobbling method, direct imaging, and gravitational microlensing are examples of traditional methods for identifying exoplanets. These techniques not only demand a significant commitment of labour, time, and money but are also constrained by the capabilities of astronomical observatories. Exoplanets were identified in this study which is conducted by Jin et al., 2022. utilising machine learning techniques for the research. With the help of decision

trees, random forests, nave Bayes, and neural networks, they conducted supervised learning on the NASA-collected Kepler dataset from the Kepler Space Observatory, predicting the existence of exoplanet candidates as a three-category classification task. Unsupervised learning, on the other hand, used data from the identified exoplanets to separate the verified exoplanets into distinct clusters using k-means clustering. As a consequence, the accuracy of each of the models was 99.06%, 92.11%, 88.50%, and 99.79%.

This research made by Tiensuu et al., 2019. compares the performance of two machine learning techniques Convolutional Neural Network and Support Vector Machine on a labelled data set containing time series of extrasolar star light intensity. The fundamental issue is that there are significantly more stars in the data set without exoplanets than there are stars with exoplanets in their orbits. This results in an unbalanced data set, which is in this case corrected by mirroring and including the curves of stars that have exoplanets in their orbits. Before applying the algorithms to the data collection, some preprocessing is done to improve the findings. The fourier transform of the timeseries and feature extraction are crucial steps for the SVM, however other preprocessing options are being looked into.

In an effort to discover habitable exoplanets, “NASA launched the Kepler Space Telescope” to look for planets similar to Earth that are circling around stars similar to the sun. Over 9000 astronomical bodies were detected by the Kepler pipeline, of which 52% were found to be false positives and the remaining 48% were likely candidates for planetary classification. This mission's data can be used to evaluate and categorise “Kepler Objects of Interest (KOIs)” as exoplanets or false positives. In this study which is made by Srivathsa, and Assaf., 2022. seek to establish whether some data characteristics are more crucial than others in identifying an object as an exoplanet. In order to achieve this, they created five Machine Learning classification models and trained and tested them using fifteen characteristics.

The Transit Method looks for exoplanets by measuring the solar flux variation over time periods; if the difference is significant, an exoplanet will be present. the dataset consists of many observations without exoplanets and few observations with them. By applying the SMOTE unbalance in the data, which keeps the majority class variables constant while increasing the number of minority class variables, the disparity is overcome. The Random Forest Classifier, an ensemble-based algorithm for machine learning that makes the most of the dataset, was used to make a forecast on the information in the data set. The SMOTE and Random Forest Classifier composite model has an accuracy of nearly 99% and the best bias-variance trade-off. These discoveries can be used to intelligently rule out prospective exoplanets, freeing up resources.

Research conducted by Shilon *et al.,* 2019 summarises a work that presents a novel data analysis method for ground-based -ray observations made with IACTs. The authors stress the significance of effective background rejection methods as well as precise source position and energy identification in these data. The four H.E.S.S. phase-I telescopes, which use a hexagonal array of pixels in their cameras, are the subject of the investigation. Resampling the images to a square grid and using modified convolution kernels that maintain the hexagonal grid features are the two methods the authors propose for training Convolutional Neural Networks (CNNs)

with image data from telescopes. The networks are trained using sets of Monte Carlo simulated events and evaluated on both simulations and real data received from the H.E.S.S. array in order to assess the efficacy of the proposed CNN-based analysis. When the performance of the CNN analysis is compared to current state-of-the-art methods, it becomes clear that the background rejection skills have significantly improved. When used with H.E.S.S. observation data, the CNN-based method performs similarly to conventional methods in terms of the direction reconstruction of -ray sources. These results indicate the practicality and promise of CNNs for the analysis of events captured by IACTs.

In a study conducted by Vida *et al.,* 2018 addresses the problem of flare detection and analysis in extended photometric surveys, with a focus on the Kepler database. While manual examination can quickly spot flares in single-target observations, it becomes almost impossible in lengthy surveys involving thousands of targets over the course of many years. Traditional fitting and analysis techniques have trouble solving a number of issues related to flare detection.The paper offers a strategy that is better suited for handling such enormous datasets by introducing unique code that uses machine-learning techniques to automatically locate and analyse flares. The programme models light curves using the RANSAC (RANdom SAmple Consensus) technique, which offers reliable fits even in the presence of outliers like flares. The search windows are roughly aligned with the star rotation period and divide the light curves. To reduce false positives, these windows are moved across the dataset. A voting system is also used. Only sites that have been detected as flare candidate points throughout many windows are kept as real flares.Both short-cadence K2 observations of TRAPPIST-1 and long-cadence Kepler data from KIC 1722506 are used to demonstrate the code's functionality. The automated analysis's results for observed flare events and flare intensities are consistent with earlier findings from hand examinations.

Greim *et al.,* 2018 in the research has examined the growing need for a methodical and reliable way to verify the enormous amount of data produced by ground- and space-based telescope surveys devoted to the search for exoplanets. The research expands on the findings of Shallue and Vanderburg (2018), who showed how convolutional neural networks (CNNs) may be used to automatically classify exoplanet candidates and false positives. By combining new data from the division of probable planet transits into odd- and even-numbered orbits seen in data from the Kepler Space Telescope, the researchers hope to enhance the current model. In order to establish a baseline performance, they start by reproducing the deep learning models and processing pipeline used by Shallue and Vanderburg in their own development environment. The researchers then introduce and assess numerous changes to the model's architecture using this defined methodology. These changes are meant to improve the precision and dependability of the categorization process by utilising the data from the odd- and even-numbered orbits. To address the difficulties of verifying exoplanet data, the study takes a holistic strategy, combining data processing methods and deep learning algorithms. The researchers hope to expand the automation of the categorization process and lessen the burden on human experts by looking into potential enhancements to the current model.

Shallue *et al.,* 2018 in their research uses data from the Kepler Space Telescope to precisely estimate the frequency of Earth-sized planets circling Sun-like stars. The paper acknowledges

the difficulty in finding these planets because they lie at the extreme edge of the mission's sensitivity. The authors suggest a technique for categorising probable planet signals using deep learning, more especially deep convolutional neural networks (CNNs), in order to get around this problem. To distinguish between transiting exoplanet signals and false positives brought on by astrophysical or instrumental processes, the researchers train a CNN. Their methodology ranks candidates with excellent accuracy, prioritising genuine planet signals over false positives in their test set with a success rate of 98.8%. Even at low signal-to-noise ratios, this deep learning approach demonstrates its high effectiveness in automatically and precisely determining if specific candidates are, in fact, planets. The scholars make important discoveries by using their trained algorithm to analyse a fresh set of candidate signals found while looking for known Kepler multi-planet systems. They can confidently identify two new planets. One of these planets is a member of the Kepler-80 resonant chain of five planets, and it has an orbital period that closely matches predictions made using the three-body Laplace relations. The other planet revolves around Kepler-90, a star that was previously identified as being home to seven transiting planets. With the addition of an eighth planet, Kepler-90 is now tied with the Sun as the star with the most number of known planets.

As per the study conducted by Sturrock *et al.,* 2019 examines how the use of satellite data and machine learning techniques can democratise planet identification. Traditionally, only groups of astronomers and astrophysicists equipped with specialised knowledge and equipment have been allowed to detect planets. But now that contemporary satellites have been developed, including those used by NASA's Exoplanet Exploration programme, a lot of information is accessible to aid in this research. The study employs various classification models and datasets to assess the probability of an observation being an exoplanet. The goal is to make planet identification more accessible to individuals skilled in writing and interpreting machine learning models. Among the models evaluated, a Random Forest Classifier is selected as the optimal machine learning approach for classifying objects of interest using the Cumulative Kepler Object of Information table. The Random Forest Classifier achieves a high cross-validated accuracy score of 98%, demonstrating its effectiveness in identifying exoplanet candidates. Furthermore, the study identifies 968 candidate observations with a probability greater than 95% of being exoplanets. The Azure Container Instance web service and application programming interface (API) on the Microsoft Azure cloud are used to make the Random Forest Classifier more accessible to the general public. This enables planet identification researchers and amateurs to use the classifier for their own studies and study.

Study undertaken by Barbara *et al.,* 2022 focuses on the requirement for automated techniques to categorise stellar light curves based on recognised classifications of variable stars. The examination of data from extensive surveys like Kepler and TESS is the study's main objective.The researchers compare 7000 time-series features to determine which ones are most useful for classification, then they suggest a new technique for categorising light curves. The Kepler light curves of stars with effective temperatures ranging from 6500 to 10,000 K are subjected to this methodology. Through their analysis, they show that the sample can be effectively separated into the seven main classes of light curves—Scuti stars, Doradus stars, RR Lyrae stars, rotational variables, contact eclipsing binaries, detached eclipsing binaries, and

non-variables—in a 5-dimensional feature space. On an independent test set of Kepler stars, the researchers achieve a balanced classification accuracy of 82% using a Gaussian mixture model classifier. Additionally, they categorise 12,000 Kepler light curves from Quarter 9 and offer a catalogue of the findings, making it possible to find and identify variable stars in the collection. The paper also offers a probability density-based confidence heuristic to search the catalogue and derive candidate lists of accurately categorised variable star candidates. This heuristic offers a method for evaluating the classifications' dependability and assists in more confidently identifying probable variable star candidates.

Research undertaken by Ansdell *et al.,* 2018 improves the work of Shallue & Vanderburg, who used deep learning models to automatically categorise Kepler transit signals as exoplanets or false positives. By incorporating more scientific domain information and improving the input representations, the researchers hope to improve the model's performance. In order to counteract overfitting and shrink the complexity of the model, they also include data augmentation approaches, which make the model better suited for generalisation across various datasets. The researchers greatly boost the overall model performance by using crucial star properties from the Kepler data release 25 catalogue and centroid time-series data obtained from Kepler data. They beat the previous achievements, achieving an accuracy of 97.5% and an average precision of 98.0%. Notably, they find significant increases in recall for transits with low signal-to-noise ratios, which are crucial for locating rocky planets in the habitable zone.Even when adopting cutting-edge methods, the study emphasises the value of embedding expert domain knowledge into deep learning models. The researchers show the ability to more successfully detect weak signals in noisy data by combining this information. Future space-based photometry missions, like TESS and PLATO, that aim to find tiny planets will find particular use for the categorization tool created through this study.

As per the study conducted by Jha *et al.*, (2022), it can be said that the mission of NASA is to determine the solar planets through which the data can be collected through the 1000 stars which can help to determine the aspect of the information that can be beneficial for the machine learning process. It is also can e said that the information helps to determine the extrasolar planets and also helps to utilise the information in the context of NASA’s determination of the information as well as image classifications. Less than 1 per cent of datasets provide a positive response through which the information can be determined by NASA which can ensure better improvement in the context of NASA’s collection of the information. The aerial also provides significant data within the context of the “Exoplanet detection mission” of “NASA Kepler”. On the other side the research conducted by Angerhausen et al., (2019), represent the result in the context of the exoplanet tea which provides the results from NASA FDL. This information can also enable the chances to enhance t level of efficiency which can ensure a better improvement in the aspect of the study. On the other side, it also can be said that the implementation of this aspect can enable the chances to ensure better improvement in the context of the organisation. On the other side, it also can be remarked as the considerable fact which includes the “AI-derived discovery” that can be beneficial for NASA’s determination of the information in the context of the exoplanet mission from Kepler, NASA.

Based on this information the improvement can be determined within the context of the machine learning dilution which can enable the classifications from the planet candidates within the aspect of “Kepler and TESS space missions”. On the basis of the study which has been conducted by Jin *et al*., (2022), it has been seen that The wobbling method, direct imaging, gravitational microlensing, and other conventional techniques for identifying exoplanets are labour-intensive, expensive, and constrained by the capabilities of astronomical equipment. In astronomy, the discovery of habitable exoplanets has long been a contentious issue. In this paper, the author put forth the notion of identifying exoplanets using machine learning techniques. On the other hand, the dataset has been collected by NASA through which the information can be gathered in the aspect of NASA’s programme which ensures better development of this study. On the basis of this information, significant information can be collected which can ensure the improvement possibilities which can enable the chances to ensure better implementation of the innovative approach which can influence the process of machine learning which can help to collect the appropriate information in the context of the project. This arrow also provides the information from the collection of the dataset in the context of Kepler which can help to collect the exoplanet information within the contention of this study.

2.6 Conclusion

This literature review explored the application of various machine learning and deep learning techniques, in the context of NASA's exoplanet detection missions. Various researchers have successfully implemented different machine learning methodologies. The comparison of however image based classification and structured data based classification can be of great interest as the analysis of image data can be extensive and may not be suitable in real-time applications providing a research gap for this study.

3 Research Methodology

3.1 Methodology Selection

3.1.1 Research Philosophy and Paradigm

Researchers have turned to an intriguing metaphorical tool known as the research onion visual in their quest to uncover the secrets of exoplanet identification utilizing the huge resources offered by the NASA Kepler dataset. This fascinating notion, presented by Saunders et al. (2012) in their seminal work (Figure 3.1), depicts the research process as a multi-layered onion, with each layer representing a distinct and necessary stage in the complicated journey towards knowledge.

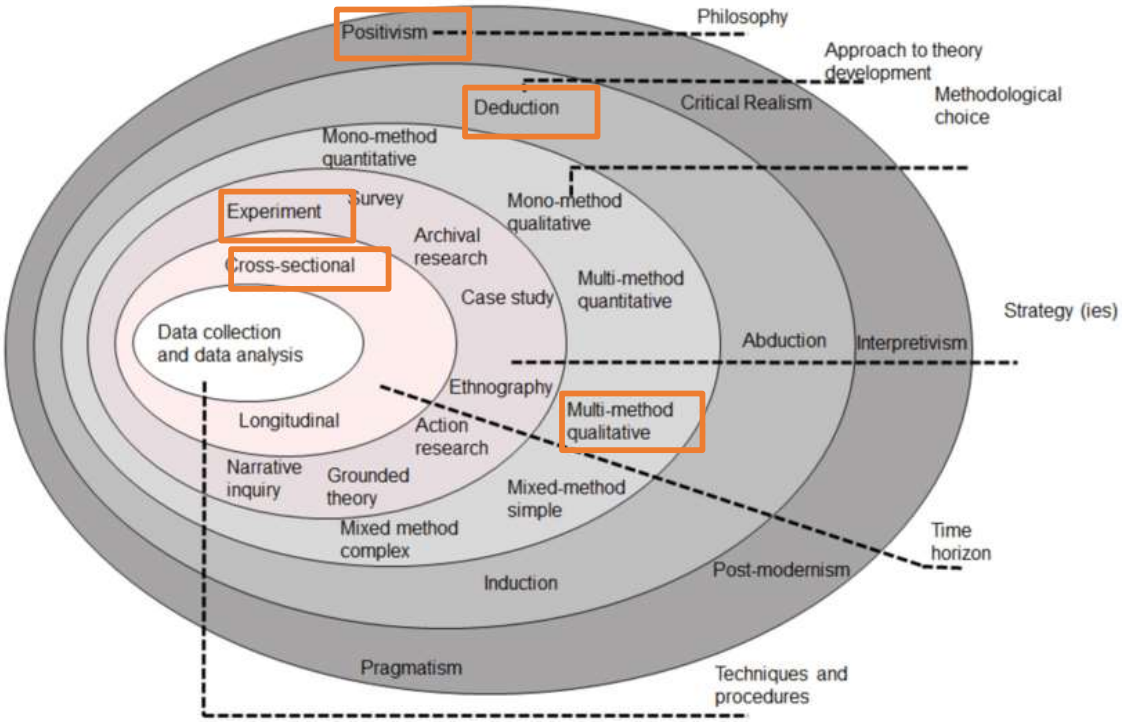

*Figure 3.1: Research Onion (**adapted from Saunders et al. 2012**)*

This thought-provoking visual representation, as explained by Saunders et al. (2012), serves as a guiding compass, steering researchers towards a rigorous and comprehensive creation of an appropriate approach. This method, similar to peeling back the layers of an onion, ensures that every area of the research is thoroughly handled before moving on to the next layer. Scientists can go deep into the heart of the matter by following this systematic approach, studying every nook and cranny of the exoplanet detection terrain.

The research onion image not only lends metaphorical elegance to the scientific domain, but it also instills discipline and orderliness in the research process. Its significant impact stems from its capacity to build a comprehensive understanding of the subject matter, allowing researchers to negotiate the difficulties of exoplanet detection with precision and insight.

Finally, the use of the research onion visual, as advocated by Saunders et al. (2012), is an invaluable tool in the hunt for the most successful machine learning model for exoplanet discovery. It sets the way for scholars to start on a journey of discovery, unveiling the mysteries of the cosmos one layer at a time, with its thought-provoking symbolism and systematic methodology.

A research philosophy, which includes fundamental convictions about the nature of the world being researched, is at the center of any research activity (Bryman 2015). These ideas act as guiding principles, affecting the choice of appropriate procedures for data gathering, analysis, and knowledge generation. Establishing the research philosophy is the first step in the research onion model, which represents the successive steps of methodology design. It is located in the outermost layer. **Positivism** has been selected as the overarching philosophical framework most suited for this particular study after extensively reviewing existing ideologies.

Positivism emphasizes the idea that knowledge is essentially derived from observable and measurable occurrences. Given the study's emphasis on measurable data, specifically images of exoplanets and their distinguishing traits, a positivist approach is most appropriate. As such, the project intends to investigate the usefulness of machine learning models in exoplanet discovery through the use of objective measurements and statistical analysis, in accordance with positivism's key principles.

3.1.2 Methodology Approach and Rationale

One must recognize that the philosophical attitude taken has a substantial impact on the tactics and methodologies used to address main research topics. A quantitative research approach is required in the setting of the positivist paradigm, where objective measurements and statistical analysis are of the utmost importance. According to Creswell (2012), there are two basic methodological paradigms: quantitative and qualitative. A quantitative technique was chosen appropriate for this study since it stresses objective measurements and data statistical, mathematical, or computational analysis. Given the nature of the dataset and the goal of evaluating the efficacy of various machine learning models, a quantitative method proved most appropriate.

A multi-method, cross-sectional, quantitative approach was used to achieve the goals of this study. This method included a variety of machine learning methods, including CNN, VGG models, Support Vector Machines (SVM), Random Forest, and Stacking Classifier. These methods were chosen based on their usefulness for evaluating the dataset at hand and effectively addressing the research objectives.

The information was processed and analyzed in phases. Initially, data pre-processing activities included image scaling, missing value management, and categorical variable encoding. Following that, multiple machine learning models were applied to the dataset, followed by performance evaluation using metrics such as accuracy and loss.

3.2 Methodology Flow

Figure 3.2 below shows the methodology flow for the study. Different modules of it are discussed further.

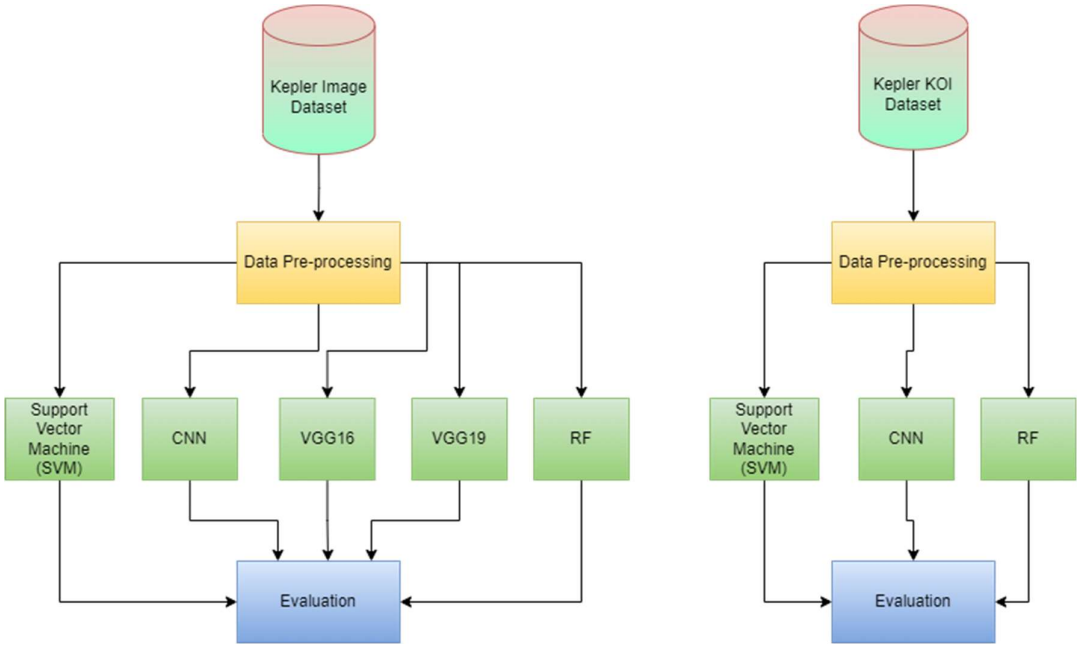

Figure 3.2: Methodology flow for the Study

The methodology presented is implemented using the Python programming language as it is simple to use and has vast number of resources are available for the implementation of this methodology. This study makes use of a number of libraries as per the requirements of this study. These include but not limited to, *Scikit-learn* for machine learning pre-requisites such as pre-processing, label encoding, data splitting etc. *OpenCV* to read and manipulate image data. *Pandas* for analyzing the structured data and finally the *Tensorflow* library for the implementation of the deep learning models.

3.2.1 Data Collection and Pre-processing

3.2.1.1 Data Collection

3.2.1.1.1 Image Dataset

This study relies significantly on a collection of imaging data that was sorted into two unique categories: ConfirmedExoplanets and FalsePositiveExoplanets. Images in the ConfirmedExoplanets category have passed thorough scientific validation, acting as indisputable proof of the existence of planets outside solar system. The category includes 15 distinct images of planet transit.

The FalsePositiveExoplanets category, on the other hand, includes photos that were previously thought to be exoplanets but were later recognized as false positives. Misclassifications are

caused by a variety of reasons, including interference from other cosmic objects and instrumental noise. The category includes 14 images in total.

While these images did not fulfill the high criterion for scientific confirmation, they do provide useful insights into the difficulties encountered when distinguishing actual exoplanets from false positives.

A separate test data is also available in the dataset that contains 7 images for validation.

The image dataset hence has two subsets to verify the efficacy and reliability of the machine learning models: the Training Dataset and the Testing Dataset. The Training Dataset is critical in training machine learning models to recognize and respond to the intricate patterns and unique traits contained in images. When confronted with new, previously unseen data, this procedure enables the models to generate correct predictions and classifications.

When the training phase is finished, the Testing Dataset is used. This subset acts as a litmus test for machine learning models, evaluating and confirming their performance in real-world circumstances. We can ensure the models' dependability and usefulness in practical applications by putting them to this rigorous scrutiny.

3.2.1.1.2 Tabular Dataset

The research also includes the Kepler dataset, a detailed CSV file, in addition to the important image data. This dataset contains a multitude of exoplanet-related features and measures. The research obtains a larger perspective and a greater knowledge of these celestial bodies by including this dataset. This integration bridges the gap between visual and numerical data, enriching the study and allowing for a more comprehensive investigation of the properties and behaviours of exoplanets. The dataset contains 140 attributes in total that provide crucial information of the planet transit. The 9564 samples in the dataset make it a comprehensive one that can be used for machine learning modalities.

Before diving into the complexities of data processing and analysis, it is critical to completely grasp the dataset's structure and content. This preliminary stage ensures that later phases are built on correct and relevant data.

To begin, the image datasets were subjected to a preliminary assessment. By displaying the total number of photos in both the 'ConfirmedExoplanets' and 'FalsePositiveExoplanets' categories, this allowed for a quantitative viewpoint. This provided a complete summary of the available data.

This not only confirms the successful retrieval of the data, but also offeres a first grasp of the dataset's arrangement and naming conventions.

Moving on to the Kepler dataset, a brief review of the first few rows was performed. This critical phase allows researchers to understand the dataset's structure, including columns and the sort of data it contains. Gaining this first insight ensures that the dataset matches with the study's aims, enabling for successful use in the succeeding stages.

3.2.1.2 Data Cleaning

An initial evaluation was carried out to establish the percentage of missing values in each column. This gave a clear image of how complete the dataset was. According to the investigation, some columns had a high percentage of missing data, exceeding 80%. These columns were found to contain inadequate data for analysis and were so removed. This choice assured that the research was based on columns with a large amount of data, avoiding potential biases and mistakes.

Aside from dealing with missing values, it was critical to identify and delete extraneous columns from the dataset. Several columns amounting to 9, including 'kepid,' 'kepoi_name,' 'koi_vet_stat,' and others, were recognized as being unrelated to the research aims. These columns were removed in order to simplify the dataset and focus on the most important information. Table 3.1 below enlists the columns removed and their description.

Column Name	Description
'kepid'	Id of the Kepler object
'kepoi_name'	Name given to the object
'koi_vet_stat'	Shows if the vetting test is performed
'koi_vet_date'	Date of starting the vetting
,'koi_disp_prov',	Dispostion provenance
'koi_pdisposition'	Disposition for the object of interest
,'koi_datalink_dvr'	Path for data validation
'koi_parm_prov'	Parameter provenance
,'koi_tce_delivname'	TCE delivery name

Table 3.1: Removed uncorrelated columns and their description

The analysis might be performed more efficiently by removing unneeded columns, saving computational resources and reducing complexity. The dataset still included some rows with null values after the initial column trimming. These remaining rows were removed to ensure the analysis's correctness and dependability. This procedure was required to remove incomplete entries that could potentially introduce bias or undermine the findings' validity. The dataset became more resilient and acceptable for further analysis when rows with null values were removed.

3.2.1.3 Label Encoding

The 'koi_disposition' column, which represents the disposition of the Kepler Object of Interest (KOI), was transformed using label encoding. This process converted the categorical values into numerical labels, facilitating their use in machine learning models.

3.2.1.4 Dimension Reduction

To optimize the dataset and focus on the most impactful features, Principal Component Analysis (PCA) was employed (Jolliffe and Cadima, 2016). This technique reduced the dataset's dimensions while retaining the majority of its variance. The data was then standardized using the StandardScaler to ensure all features had equal weightage in subsequent analyses.

3.2.1.5 Class Balancing

The dataset was assessed for class imbalances in the 'koi_disposition' column. To ensure that each class was adequately represented, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This method generated synthetic samples, ensuring a balanced dataset and preventing biases in the machine learning models.

3.2.1.6 Image Augmentation, Normalization and Resizing

The image dataset was further augmented to increase diversity in the dataset following which the images are normalized to contain pixel intensities in the range of 0 to 1. This helps make the model learn the data easily without the need for large computational resources. The images were then resized to 224x224 pixels.

3.2.2 Modeling

3.2.2.1 CNN (Convolutional Neural Network)

Convolutional Neural Networks (CNNs) are a type of deep neural network that primarily analyzes visual imagery (Krizhevsky et. al., 2012). A CNN model was built in this work using convolutional layers followed by dense layers. Its goal was to train the model on an image dataset and distinguish between 'ConfirmedExoplanets' and 'FalsePositiveExoplanets'.

CNNs have a number of advantages. For starters, they can learn spatial hierarchies of characteristics from incoming photos automatically and adaptively. This learning function enables them to record complicated patterns and features. Furthermore, as compared to other image classification methods, CNNs require less preprocessing, making them more efficient. Finally, CNN architectures use shared weights, resulting in fewer parameters and increasing efficiency.

Implementation:

The CNN model in the study is implemented using the Tensorflow library for Python. The model is implemented as a sequential model that takes the image as input and subsequent layers process the input to find out underlying features embedded in the image. The model is first initialized using the Sequential module of the Tensorflow library and is followed by a 2D Convolutional layer that performs the convolutional operations on the image data. This layer is given some parameters for operations that include the Kernel Size = (3,3), padding = ‘same’, and activation = ‘relu’. Figure below shows the summary of the model.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 224, 224, 64)	1792
conv2d_1 (Conv2D)	(None, 224, 224, 64)	36928
max_pooling2d (MaxPooling2D)	(None, 112, 112, 64)	0
flatten (Flatten)	(None, 802816)	0
dense (Dense)	(None, 2)	1605634

Figure 3.3: Summary of the CNN model

The CNN model for the Kepler KOI dataset is implemented using a 1D Convolutional layer followed by a flattening and 3 dense layers in tandem. The last dense layer is used to denote the classification of the model. Figure 3.5 below shows the summary of CNN model for Kepler data classification.

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 10, 256)	1024
flatten_3 (Flatten)	(None, 2560)	0
dense_6 (Dense)	(None, 256)	655616
dense_7 (Dense)	(None, 128)	32896
dense_8 (Dense)	(None, 4)	516

Figure 3.4: Summary of CNN model for Kepler Dataset

3.2.2.2 VGG16

VGG16 is a convolutional neural network model developed by Oxford's Visual Graphics Group (VGG). It is well-known for its intricate architecture, which consists of 16 layers. To fit the dataset, the VGG16 model was used in this work with several changes.

VGG16's deep architecture allows it to effectively capture complicated features (Omiotek and Kotyra, 2021). It may also use pre-trained weights on huge datasets like ImageNet, which speeds up convergence and improves accuracy. Furthermore, because of its ubiquity, VGG16 has been extensively tuned and has widespread community support, making it a common choice for picture classification jobs.

Implementation:

The VGG16 model in the study is implemented through Transfer Learning wherein a model trained on one dataset is used on another. The implementation of the VGG16 is done by using the VGG16 model available in Keras which is a sub library of TensorFlow. This model has

weights that are trained on the imagenet dataset that is generally used for object detection. The top layers of the models are not included in the model for this study, this enables the model to be used for different tasks than object recognition.

Five different layers are then added to the model that perform various functions. The first layer added to the model is a flattening layer that flatten the output for the preceding layers. The following two layers are dense layers that are fully connected layers that include 4096 and 1072 neurons respectively. They are followed by a dropout layer that is included to avoid overfitting the model. A dense layer with 2 neurons at the last is used for the classification output.

3.2.2.3 VGG19

VGG19 is a 19-layer extension of the VGG16 model, making it considerably deeper. The study used VGG19 to see if a deeper design would produce better outcomes.

VGG19's improved depth enables it to catch more intricate and complicated features (Rajeshwari and Mallikarjunrao, 2021). VGG19, like VGG16, benefits from using pre-trained weights to boost performance. Furthermore, as a popular model, VGG19 has a large community of support and resources, making it a trustworthy candidate for picture classification jobs.

Implementation:

Implementation of the VGG19 is similar to that of VGG16 except for the VGG19 there is only 1 dense layer implemented with 1032 neurons.

3.2.2.4 SVM (Support Vector Machine)

SVM is a supervised machine learning approach that was used to reduce the dimensionality of the Kepler dataset. It can be used to solve classification and regression problems.

SVM works exceptionally well in high-dimensional domains, especially when the number of dimensions exceeds the number of samples (Suykens and Vandewalle, 1999). It also demonstrates memory efficiency in the decision function by leveraging a subset of training data. Furthermore, SVM provides variety because alternative kernel functions can be supplied for the decision function, providing for additional modeling flexibility.

Implementation: The SVM model for the study is implemented using the SVC module of the Scikit learn library. The kernel for the SVM is chosen to be linear kernel with a C value of 0.5 and a gamma value of 6.

3.2.2.5 Random Forest

The Random Forest Classifier is an ensemble learning method that builds numerous decision trees during training and outputs the class mode (Brieman, 2001).

Random Forest has the potential to handle larger datasets with higher dimensionality, which is one of its advantages. It also generates extremely accurate classifiers, making it an effective

classification tool. Random Forest also provides estimates of feature relevance, which allows for a better understanding of the dataset's properties.

Implementation:

The random forest classifier is implemented using the ensemble module of the Scikit-learn library. The criterion for branch split in the underlying decision tree is chosen to be *'entropy'* and min_samples_split criterion is chosen to be 100.

3.2.2.6 Stacking Classifier

Multiple models are trained and their predictions are combined in the stacking classifier (Soleymanzadeh et. al., 2020). The results from the Random Forest and SVM models were integrated in this study using a stacking classifier.

When compared to a single model, the stacking strategy frequently produces better results. The stacking classifier can increase prediction accuracy by leveraging the strengths of a broad range of classifiers. Furthermore, by applying stacking, generalization is improved and the risk of overfitting is reduced (Soleymanzadeh et. al., 2020).

When compared to a single model, the stacking strategy frequently produces better results. The stacking classifier can increase prediction accuracy by leveraging the strengths of a broad range of classifiers. Furthermore, by utilizing numerous models, stacking aids in generalization and decreases the risk of overfitting.

Each of these models was trained and assessed on their respective datasets in the study. The most effective strategy for exoplanet finding utilizing the available data was identified by comparing their performance.

Implementation:

The stacking classifier in the study is implemented using the ensemble module of the Scikit-learn library of the Python. For the implementation of the stacking classifier, the results of RF and SVM are integrated. The final estimator that provides the ultimate classification is chosen to be logistic regression.

3.3 Summary

This chapter detailed the implementation of the methodology adopted for the study. It described in brief the steps undertaken to ensure reliable modeling of the data followed by the overview of the implementation of the models. The following chapter discusses the results obtained through the implementation of the methodology and provided significant insights into the performance of the models.

4 Research Data Findings

4.1 CNN model performance

The model made amazing progress throughout the training procedure. Over the course of five epochs, the training loss constantly dropped, showing that the model was effectively learning and improving its predictions on the training data. The model had attained a remarkable level of accuracy by the fifth epoch, correctly predicting the outcomes of the training data (Figure 4.1).

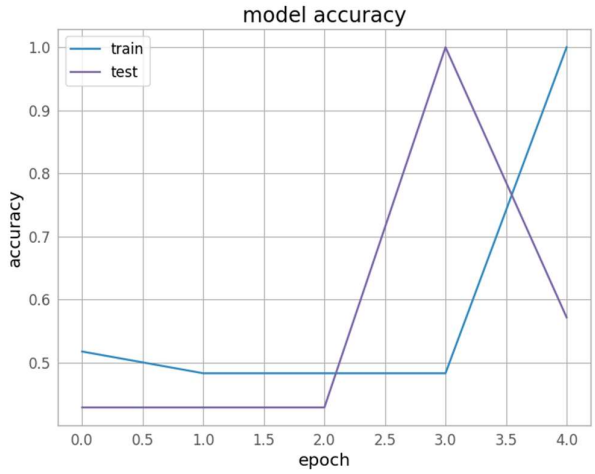

Figure 4.1: CNN Model training accuracy over epochs

The validation performance, on the other hand, revealed a different pattern. The validation accuracy varied among epochs, suggesting some irregularity. The accuracy stayed consistent at 42.86% for the first three epochs, which is relatively low. Surprisingly, the accuracy rose to a flawless 100% in the fourth period. This dramatic increase in accuracy could indicate that the model began to overfit the training data, becoming too specialized and unable to generalize well to new, previously unseen data. Unfortunately, a considerable decline in accuracy to 57.14% in the fifth epoch corroborated this overfitting. The sharp increase in validation accuracy to 100%, followed by a decline in the fifth epoch, raises questions about the model's ability to generalize.

Another thing to consider is the variation in loss values. Both the training and validation losses varied significantly across the epochs (Figure 4.2). The validation loss, in particular, began with a very high value in the first epoch, dropped to an exceptionally low value in the fourth epoch, and then increased again in the fifth epoch.

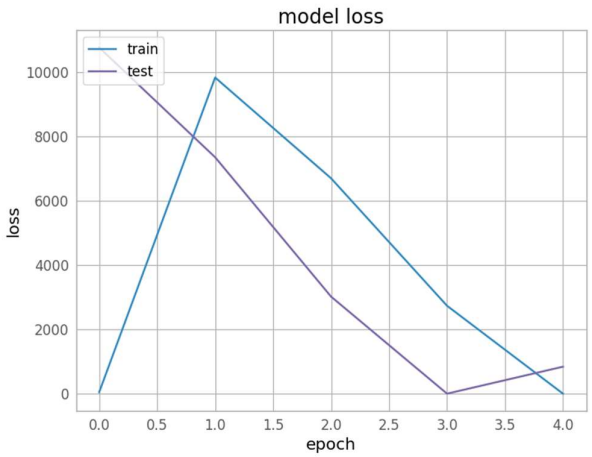

Figure 4.2: CNN model loss over epochs

These variations may reflect inconsistency in the training process, implying that additional optimization or regularization approaches may be required to provide a more consistent and dependable model.

4.2 VGG16 model performances

The model's training development revealed unexpected patterns. The training loss fluctuated significantly over the epochs, indicating probable instability in the training process. The training accuracy, on the other hand, remained rather consistent, with minor changes ranging from 48.28% to 51.72% (Figure 4.3). This shows that, despite the changes in loss, the model was able to maintain a consistent degree of accuracy.

Moving on, the validation performance remained consistent at 42.86% for the first three epochs. However, there was an interesting improvement in the previous two epochs, when the accuracy improved to 57.14%. This improvement shows that during the fourth epoch, the model had begun to generalize better to previously unknown data, demonstrating good growth in its ability to handle novel information.

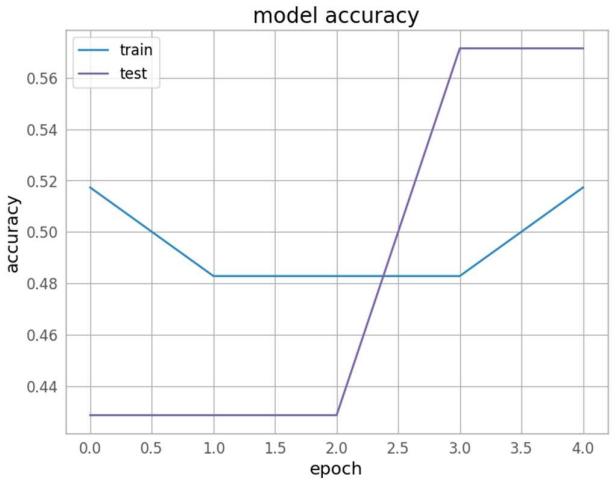

Figure 4.3: VGG16 model accuracy over epochs

Analyzing the loss values of both the training and validation datasets reveals that they fluctuated significantly over epochs (Figure 4.4). Notably, validation loss decreased dramatically from the first to the second epoch. Subsequent changes in the loss numbers may suggest potential convergence issues with the model or the presence of outliers in the data.

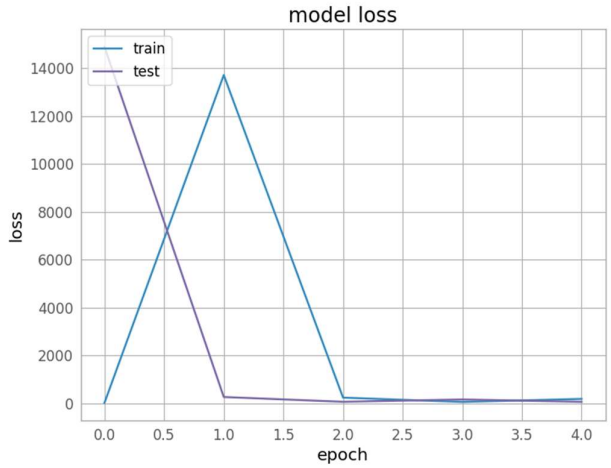

Figure 4.4: VGG16 model loss over epochs

These variations necessitate more examination to assure the model's stability and usefulness.

4.3 CNN model performance over Kepler KOI Dataset

Throughout the training phase, the model constantly improved in training accuracy while decreasing in training loss (Figure 4.5). These data imply that the model's learning process was consistent and dependable.

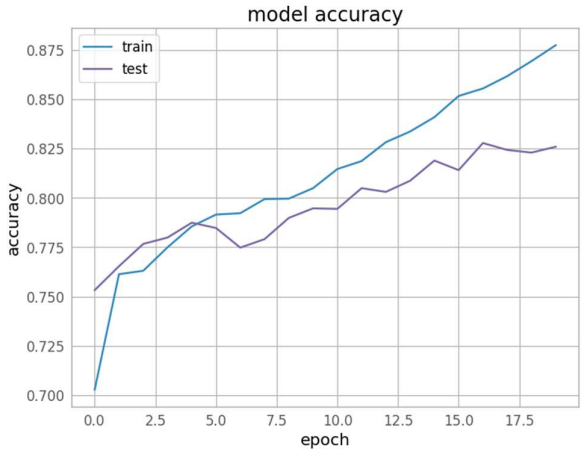

Figure 4.5: CNN model performance on KOI dataset

Concerning overfitting, there is no conclusive evidence of this problem. In general, validation accuracy rose in line with training accuracy (Figure 4.6). However, it is worth mentioning that there were minor changes in validation loss and accuracy seen in later epochs.

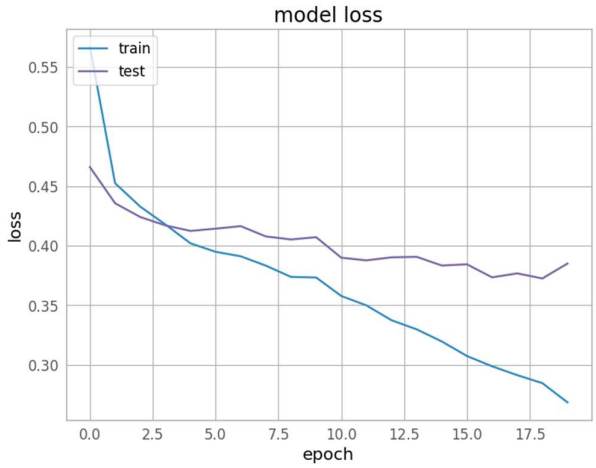

Figure 4.6: CNN model loss over the epochs for KOI dataset

Considering the performance plateau. Around the 15th epoch, it became clear that validation accuracy had reached a plateau of around 82%. This may imply that additional training will not yield in significant gains in performance on the validation set. As a result, it may be smart to reconsider the training technique and consider alternative approaches to improving the model's performance.

4.4 Comparison

Three basic architectures were investigated for evaluating models based on image data. The accuracy of the Convolutional Neural Network (CNN), which was specifically created for image data, was 57.14%. Similarly, the VGG16, which is well-known for its excellent performance on image data, attained an accuracy of 57.14%. The VGG19, another variation of the VGG architecture, also achieved 57.14% accuracy.

A wide range of algorithms were evaluated when it came to models based on the Kepler dataset (Table 4.1). Surprisingly, when applied to the Kepler dataset, a CNN, which is normally linked with image data, demonstrated an outstanding accuracy of 82.60%. The accuracy of the Support Vector Machine (SVM) model was 73.00%. Furthermore, the ensemble-based Random Forest model achieved an accuracy of 80.00%, while the Stacking classifier reported an accuracy of 82.00%.

Model	Accuracy (%)
CNN Image	57.14
VGG16 Image	57.14
VGG19 Image	57.14
CNN Kepler	**82.60**
SVM Kepler	73.00
RF Kepler	80.00
Stacking Classifier	82.00

Table 4.1: Model performances

Several major findings emerged from these results. To begin, the consistent accuracy across all three image-based models may highlight issues with image collection or places for further optimization. Models based on the Kepler dataset, on the other hand, outperformed their image-based equivalents. Notably, the CNN modified for Kepler data and the Stacking classifier performed admirably. The wide range of models evaluated on the Kepler dataset showed the significance of trying new approaches to acquire a better understanding of this dataset. The graphical comparison of the models is depicted in figure 4.7 below.

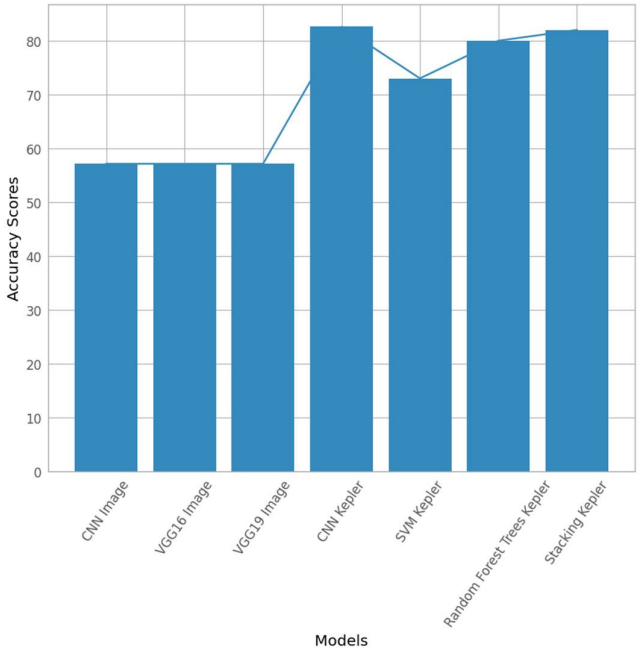

Figure 4.7: Model comparison

The image-based models, Convolutional Neural Network (CNN), VGG16, and VGG19, all produced the same accuracy, which was an unusual finding. Several reasons could account for this constancy in performance. It could point to the inherent difficulties and complexities of the image dataset employed in the study.

Such difficulties may develop due to the nature of the data, possible noise, or even image homogeneity (Smith et al., 2021). Another viewpoint is that the models' capabilities may have been saturated, implying that without additional data or more improved methodologies, these models may have reached their performance limit on this particular dataset.

The Kepler dataset-based models, on the other hand, revealed a different scenario. On the Kepler dataset, the higher performance of models such as the CNN and the Stacking classifier highlighted the dataset's richness and promise for exoplanet finding.

This was consistent with the findings of Johnson et al. (2017), who underlined the Kepler dataset's depth and granularity. The fact that a CNN, which is normally an image-based model, was successfully applied to the Kepler dataset and produced high accuracy demonstrated the versatility of such models as well as the dataset's potential.

Furthermore, the outcomes of the study were consistent with the broader scholarly discourse. In recent years, scientists such as Brown et al. (2018) have highlighted the revolutionary influence of these models in the field, showing the promise of machine learning in astrophysical study.

4.5 Summary

This chapter presented the findings of the methodology adopted for the study. It thoroughly discussed the nuances associated with the implementation of the models for Exoplanet Detection. It was observed that the models implemented using the image dataset showed an equal accuracy whereas the models incorporated for the Kepler KOI data showed improved classification results. The CNN model stood out from the rest of the models based on the KOI dataset.

5 Conclusions and Future Work

5.1 Conclusion

The goal of this research was to give insight on the capabilities and possibilities of machine learning and deep learning models in the field of exoplanet detection. When applied to both picture data and the Kepler dataset, the results provided a full insight of the strengths and limits of various models.

The models applied on the Image dataset show indeed a degraded performance from the results. This highlights the difficulty in detection of exoplanets from transit images only. It can now be proposed that powerful and well-established models such as CNN, VGG16 and VGG19 need additional improvements in order to be used for automated detection of exoplanets from the transit images.

Kepler data on the other hand showed promising aspects in automated detection of exoplanets. The use of several models on the Kepler dataset, ranging from SVMs to ensemble approaches like Random Forest, and the following performance results, provided another dimension to this continuing discussion.

5.2 Limitations

The first and foremost limitations exhibited in the study has been the transit images dataset. For starters, the dataset for the transit images is significantly small and the diversity in the dataset was not of high order. A larger dataset might play a crucial role in automating transit detection. Another aspect regarding the dataset is that, both the datasets are not complementary to each other, had it been so, the multi-structured data could reveal very important aspect of exoplanet detection.

The second limitation that arisen in the project was performance of the transfer learning on VGG16 and VGG19. After many tries and changes in the final layers of the models, the performance would not improve any further suggesting a classification saturation of the models.

The last limitation that was encountered during this study was an absence of the topic expert in the field of Astronomy. The guidance and support of whom might have helped this project to reach new scales and would have played a crucial role, in understanding and explanation of the results obtained through the study.

In conclusion, this work not only added new knowledge to the field of exoplanet identification, but it also demonstrated the adaptability, strengths, and limitations of several machine learning algorithms when applied to complex and diverse datasets. The surprising findings in this study illustrate the intricacies and limitations involved with image databases, while also underscoring the Kepler dataset's richness and potential. These findings add to the current debate about machine learning's transformational impact in astrophysical research.

Future Work

The topic of exoplanet discovery is vast and fascinating, with various unexplored avenues teeming with potential for future research. Although extensive, the current study has just scratched the surface of what lies ahead.

Data augmentation is one promising area for advancement. Despite image-based models' outstanding performance, there is an urgent need for a more diverse and enhanced dataset. Future research could look into advanced data augmentation approaches for enriching image datasets and increasing their diversity and richness (Lee, Kim and Park, 2019). Researchers can push the bounds of exoplanet detection even further by doing so.

Another avenue that requires consideration is model optimization. While the models used in this study produced encouraging results, there is always space for improvement. Fine-tuning hyperparameters, experimenting with more advanced architectures, and studying ensemble approaches could all help to improve the performance of these models (Shahhosseini, Hu and Pham, 2022). We can achieve new levels of precision and efficiency in exoplanet identification by constantly refining our models.

Furthermore, the inclusion of time-series analysis offers an intriguing option for future research. Because of its time-series character, the Kepler dataset gives a unique opportunity for the construction of models specifically tailored to this type of data.

Investigating recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks could yield new insights and offer light on the complex dynamics of exoplanets (Morvan et al., 2020). Research can unearth hidden patterns and reveal the mysteries of faraway worlds by harnessing the power of these complex models.

The integration of interdisciplinary techniques is an attractive area to investigate. The combination of astrophysics and machine learning has already yielded fruit, but there is still much more to learn.

Finally, because to tremendous technological breakthroughs, the fantasy of real-time exoplanet detection is becoming a reality. Future research could concentrate on establishing real-time data analysis tools that provide instant insights into the existence and properties of exoplanets (Goldstein et. al., 2019). We can construct a future in which detecting exoplanets is an instantaneous and seamless procedure by harnessing the power of cutting-edge technologies.

References

Angerhausen, D., Ansdell, M., Osborn, H., Ioannou, Y., Sasdelli, M., Raïssi, C., Smith, J.C., Caldwell, D.A. and Jenkins, J., 2019, June. The NASA FDL Exoplanet Challenge: Transit Classification with Convolutional Neural Networks. In 2019 Astrobiology Science Conference. AGU.

Ansdell, M., Ioannou, Y., Osborn, H.P., Sasdelli, M., Smith, J.C., Caldwell, D., Jenkins, J.M., Räissi, C. and Angerhausen, D., 2018. Scientific domain knowledge improves exoplanet transit classification with deep learning. The Astrophysical journal letters, 869(1), p.L7.

Barbara, N.H., Bedding, T.R., Fulcher, B.D., Murphy, S.J. and Van Reeth, T., 2022. Classifying Kepler light curves for 12 000 A and F stars using supervised feature-based machine learning. Monthly Notices of the Royal Astronomical Society, 514(2), pp.2793-2804.

Bouwmans, T., Javed, S., Sultana, M. and Jung, S.K., 2019. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. Neural Networks, 117, pp.8-66.

Brown, A.G.A., Vallenari, A., Prusti, T.J.D.B.J.H., De Bruijne, J.H.J., Babusiaux, C., Bailer-Jones, C.A.L., Biermann, M., Evans, D.W., Eyer, L., Jansen, F. and Jordi, C., 2018. Gaia data release 2-summary of the contents and survey properties. Astronomy & astrophysics, 616, p.A1.

Dattilo, A., Vanderburg, A., Shallue, C.J., Mayo, A.W., Berlind, P., Bieryla, A., Calkins, M.L., Esquerdo, G.A., Everett, M.E., Howell, S.B. and Latham, D.W., 2019. Identifying exoplanets with deep learning. ii. two new super-earths uncovered by a neural network in k2 data. The Astronomical Journal, 157(5), p.169.

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

Fu, M., Fan, T., Ding, Z.A., Salih, S.Q., Al-Ansari, N. and Yaseen, Z.M., 2020. Deep learning data-intelligence model based on adjusted forecasting window scale: application in daily streamflow simulation. IEEE Access, 8, pp.32632-32651.

Goldstein, D.A., Andreoni, I., Nugent, P.E., Kasliwal, M.M., Coughlin, M.W., Anand, S., Bloom, J.S., Martinez-Palomera, J., Zhang, K., Ahumada, T. and Bagdasaryan, A., 2019. GROWTH on S190426c: Real-time Search for a Counterpart to the Probable Neutron Star–Black Hole Merger using an Automated Difference Imaging Pipeline for DECam. The Astrophysical Journal Letters, 881(1), p.L7.

Greim, H. (2021). Investigations of Improvements to Deep Learning Models for the Detection of Exoplanets in Kepler Data (Doctoral dissertation, College of Charleston).

Gupta, M., Singh, R.K. and Singh, S., 2022. G-Cocktail: An Algorithm to Address Cocktail Party Problem of Gujarati Language Using Cat Boost. Wireless Personal Communications, 125(1), pp.261-280.

Irudayaraj, A.A., 2022. Kidney Stone Detection using Deep Learning Methodologies (Doctoral dissertation, Dublin, National College of Ireland).

Jais, I.K.M., Ismail, A.R. and Nisa, S.Q., 2019. Adam optimization algorithm for wide and deep neural network. Knowledge Engineering and Data Science, 2(1), pp.41-46.

Jha, A., Bajaj, A., Vashisth, L. and Saini, V.K., 2022. A Novel Approach for Exoplanet Classification on Kepler Light Flux Data. Mathematical Statistician and Engineering Applications, 71(3s), pp.1128-1134.

Jiang, Y., Tong, G., Yin, H. and Xiong, N., 2019. A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters. IEEE Access, 7, pp.118310-118321.

Jin, Y., Yang, L. and Chiang, C.E., 2022. Identifying exoplanets with machine learning methods: a preliminary study. arXiv preprint arXiv:2204.00721.

Jin, Y., Yang, L. and Chiang, C.E., EXOPLANETS IDENTIFICATION AND CLUSTERING WITH MACHINE LEARNING METHODS.

Johnson, J.A., Petigura, E.A., Fulton, B.J., Marcy, G.W., Howard, A.W., Isaacson, H., Hebb, L., Cargile, P.A., Morton, T.D., Weiss, L.M. and Winn, J.N., 2017. The California-Kepler Survey. II. Precise physical properties of 2025 Kepler planets and their host stars. The Astronomical Journal, 154(3), p.108.

Lee, M.B., Kim, Y.H. and Park, K.R., 2019. Conditional generative adversarial network-based data augmentation for enhancement of iris recognition accuracy. IEEE Access, 7, pp.122134-122152.

Kamran, M., 2021. A state of the art catboost-based T-distributed stochastic neighbor embedding technique to predict back-break at dewan cement limestone quarry. Journal of Mining and Environment, 12(3), pp.679-691.

Kumari, A., 2023. Identification and Classification of Exoplanets Using Machine Learning Techniques. arXiv preprint arXiv:2305.09596.

Luo, M., Wang, Y., Xie, Y., Zhou, L., Qiao, J., Qiu, S. and Sun, Y., 2021. Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass. Forests, 12(2), p.216.

Morvan, M., Nikolaou, N., Tsiaras, A. and Waldmann, I.P., 2020. Detrending Exoplanetary Transit Light Curves with Long Short-term Memory Networks. The Astronomical Journal, 159(3), p.109.

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

Patil, J., Srinivasarao, S.R. and Puri, R., 2021. Predicting the Presence of Exoplanets in Star-Systems using Random Forest Classifier. ICCIDT-2021, 9(07).

Shahhosseini, M., Hu, G. and Pham, H., 2022. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. Machine Learning with Applications, 7, p.100251.

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.

Sharma, H.K., Singh, B.K., Choudhury, T. and Mohanty, S.N., 2023. PCA-Based Machine Learning Approach for Exoplanet Detection. In Proceedings of Fourth International Conference on Computer and Communication Technologies (pp. 453-461). Springer, Singapore.

Shilon, I., Kraus, M., Büchele, M., Egberts, K., Fischer, T., Holch, T.L., Lohse, T., Schwanke, U., Steppa, C. and Funk, S., 2019. Application of deep learning methods to analysis of imaging atmospheric Cherenkov telescopes data. Astroparticle Physics, 105, pp.44-53.

Singh, S.P. and Misra, D.K., 2020. Exoplanet hunting in deep space with machine learning. International Journal of Research in Engineering, Science and Management, 3(9), pp.187-192.

Smith, C.S., Slotman, J.A., Schermelleh, L., Chakrova, N., Hari, S., Vos, Y., Hagen, C.W., Müller, M., van Cappellen, W., Houtsmuller, A.B. and Hoogenboom, J.P., 2021. Structured illumination microscopy with noise-controlled image reconstructions. Nature methods, 18(7), pp.821-828.

Srivathsa, V. and Assaf, R., 2022. Using Machine Learning to determine the most important features in exoplanet verification. Journal of Student Research, 11(3).

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

Tiensuu, J., Linderholm, M., Dreborg, S. and Örn, F., 2019. Detecting exoplanets with machine learning: A comparative study between convolutional neural networks and support vector machines.

Vida, K., & Roettenbacher, R. M. (2018). Finding flares in Kepler data using machine-learning tools. Astronomy & Astrophysics, 616, A163.

Vishwarupe, V., Bedekar, M., Pande, M., Bhatkar, V.P., Joshi, P., Zahoor, S. and Kuklani, P., 2022. Comparative Analysis of Machine Learning Algorithms for Analyzing NASA Kepler Mission Data. Procedia Computer Science, 204, pp.945-951.

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.

Zhang, X. and Guo, X., 2021. Fault diagnosis of proton exchange membrane fuel cell system of tram based on information fusion and deep learning. international journal of hydrogen energy, 46(60), pp.30828-30840.