Harnessing the Power of Convolutional Neural Network for Exoplanet Discovery

Gaurav and Sumit Gupta

Abstract-**The discovery of planets apart from Earth that can sustain lives has always been fascinating as well as challenging. Discussion around such planets, popularly termed as "Exoplanets" have been doing the rounds for quite some time now. These exoplanets are often considered to be "Earth-like" or "habitable" because they may have conditions that could potentially support life. This work focuses on how Deep Learning techniques can be useful in identifying potential exoplanets. To do so, astronomical data gathered by space telescopes such as Kepler and BRITE have been utilized. The method employed to detect exoplanets is Transit Photometry along with Convolutional Neural Network. The study highlights the limitations of small training datasets and suggests the use of data augmentation techniques to increase the size of the training dataset, and the transfer learning approach to improve the performance of the classification models. The research offers valuable insights into the nature and diversity of exoplanets and may open avenues for future discoveries. With a performance accuracy of 96.67%, the proposed approach showcases merit and hence can prove to be a harbinger in exploring planetary habitability in the colossal space.**

Index Terms- **Earth, Exoplanet, Convolutional Neural Network, Light Curve, SMOTE, Transit Photometry.**

I. Introduction

A hunt for an alternative to Earth has already started. The origin of this discussion is the challenges our planet may face in the coming decades. Environmental dangers of the twenty-first century, such as invading species, diseases, pollution, and changing climate have put human populations in danger. To address these environmental issues and protect our species and its habitats, the need of the hour is to look for other planets in our solar system that are habitable and can sustain life. Such habitable planets are quite commonly referred to as 'exoplanets'. An exoplanet is any planet in or around the solar system. Most of such planets orbit other stars, but free-floating exoplanets, called rogue planets, orbit the galactic center and are not gravitationally bound to any star.

The objective of the current work is to develop and test Deep Learning models that can identify patterns in the light curves of stars which may indicate the presence of exoplanets, especially those planets that are habitable in nature. The proposed methodology involves several crucial steps, including preprocessing of data to extract useful features, training of the Convolutional Neural Network, a well-known Deep Learning model, and evaluating model's performance using benchmark performance metrics.

II. Literature survey

In the field of exoplanet detection, identifying planetary transit candidates in TESS full-frame image light curves is an important problem. One approach to solving this problem is to use Convolutional Neural Networks (CNNs), which have been shown to be effective at analyzing spatial patterns in image data.

Numerous studies [1] have investigated the utilization of convolutional neural networks (CNNs) in the identification of exoplanets, with a specific mention of the research conducted in [2]. However, a common challenge encountered when employing CNNs for exoplanet detection is the scarcity of available training data. The research paper acknowledges that the number of instances is typically insufficient to effectively train a Neural Network on the scale employed in their study without succumbing to overfitting. This issue underscores the importance of having larger and more diverse training datasets for successful exoplanet detection. Various approaches have been suggested to tackle this problem, such as employing data augmentation techniques to artificially expand the size of the training dataset, as well as adopting transfer learning, which involves pre-training a network on a substantial dataset and subsequently fine-tuning it on a smaller dataset.

Despite the constraints posed by limited training datasets, CNNs have demonstrated encouraging outcomes in the field of exoplanet detection and remain a prominent tool in ongoing research. Nevertheless, future investigations could venture into exploring alternative approaches to CNNs, such as recurrent neural networks or hybrid models that amalgamate multiple types of neural networks. These alternative approaches aim to

Gaurav is with University Institute of Technology, The University of Burdwan, Golapbag (North), Burdwan-713104, West Bengal, India (e-mail: gaurav170101@gmail.com).

Sumit Gupta is with University Institute of Technology, The University of Burdwan, Golapbag (North), Burdwan-713104, West Bengal, India (e-mail: sumitsayshi@gmail.com).

enhance the performance of exoplanet detection systems and expand the scope of research in this domain.

Another similar study [3] investigates the use of CNNs to identify exoplanet transits in BRITE data. Despite the success of this work, the authors have highlighted the pressing limitations of their approach. As stated in this work, the use of Machine Learning techniques in exoplanet detection requires careful interpretation of the results and the need for follow-up observations to confirm potential exoplanet candidates. The CNN model used in the study was successful in identifying plausible transit candidates, but the authors stress the need for more observational data in future studies to provide firm confirmations of these candidates. The use of a small training dataset may limit the generalizability of the CNN model to other datasets, which is a potential constraint to the work. The performance of the CNN models may be enhanced in future experiments, according to the authors, by using transfer learning methods and larger, more varied training datasets.

When compared to other experimentation outputs, research studies [4] and [5] that proposed a CNN-based ensemble model for exoplanet detection and a machine learning data rejection algorithm for transiting exoplanet light curves both produced subpar results.

III. Proposed Methodology

The method employed in this proposed work is based on Transit Photometry [6]. Essentially, this method involves observing a star and its surrounding planetary system. In this approach, an exoplanet planet revolves around its star and captures the star's light intensity. When a planet passes in front of a star, a segment of the star is obscured and the intensity of the light temporarily decreases. The amount of the light intensity dip and the duration of the event allow us to assess whether or not the rotating body is an exoplanet.

As seen from the figure (Fig. 1), imagine a planet revolvingaround a star like our own sun. As it moves in its orbit, it occasionally passes between us and the star. This causes a small portion of the star's light to be blocked, resulting in a decrease in brightness that we can observe from Earth. By studying these dips in brightness, we can determine whether or not the object that is causing them is an exoplanet.

Transit Photometry is just one of the methods used by astronomers to detect exoplanets, but it has proven to be highly effective [7]. By studying these exoplanets, we can learn more about the universe and explore many planetary systems that exist beyond our own.

Convolutional Neural Networks are widely used in various fields such as Medical Imaging, Autonomous Driving, and Natural Language Processing. They are a powerful tool for Image Processing and Analysis due to their high accuracy and ability to work with large datasets. In this proposed work, light intensities are considered as pixels and spatial patterns are identified amongst them by examining the relationships between different intensities. To detect exoplanets a special type of neural network known as Convolutional Neural Network (CNN) architecture is employed so as to learn and extract the relevant features from the data.

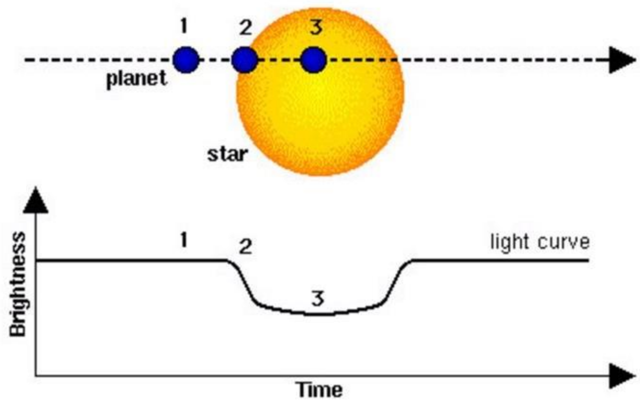

Fig. 1. Light curve of transiting exoplanet.

The CNN architecture is particularly well-suited foranalyzing image data because it uses filters to identify and extract patterns in the data. These filters are applied to small, overlapping regions of the image (or in this case, the set of light intensities), allowing the network to identify patterns and features at multiple scales. The CNN architecture can then use this information to make accurate predictions about the presence of exoplanets based on the spatial patterns identified in the light intensity data.

The proposed model implements a CNN model using the Keras API with TensorFlow as the backend. The model architecture is based on the AlexNet CNN [8] that has been used for ImageNet Classification. It comprises several layers, including convolutional layers, pooling layers, and dense layers. The convolutional layers apply filters to the input signal to extract features. The pooling layers downsize the feature maps to reduce their size and increase their robustness to variations in the input. The dense layers perform classification based on the extracted features.

The next challenge comes is to segregate exoplanets on the basis of whether there is presence of oxygen or not. Oxygen is a biomarker for 'Habitable' exoplanet. In only one of the many ionospheres in our Solar System, i.e., the Earth's, the ionised form of atomic oxygen (O+) is the dominant ion species at the altitude of maximum electron density. If oxygenic photosynthesis were not an ongoing process that continuously influences the terrestrial atmosphere, this ionospheric composition would not exist [9]. This implies that presence of O+ in the ionosphere is acting as a biomarker. Detection of O+ ions in the ionosphere can be done using extreme ultraviolet lithography.

IV. Implementation & Results

A. Dataset Description

The Exoplanet Hunting in Deep Space dataset used in this work has been downloaded from Kaggle [13]. The sample plot as seen from Fig. 2 describes how the flux (or light intensity) of thousands of stars has changed through time. There is a binary label of 2 or 1 for each star. Label 2 shows that there is evidence of at least one exoplanet orbiting the star; other observations indicate multi-planet systems.

The dataset is found to be extremely unbalanced. Only 1% of the data points are for exoplanets, while 99% are for non- exoplanets. Thus, the observations in the dataset with 3198 features each have been divided into training and validation (or test) sets. There are 5087 observations (rows) in the training set with 5050 non-exoplanets and 37 exoplanets. The test set comprises 570 observations (rows) with 5 exoplanets and 565 non-exoplanets. The training set is then over sampled using SMOTE (Synthetic Minority Over-sampling Technique), which generates Synthetic Examples that are similar to those seen in the dataset rather than duplicating Minority Class data points. There is no possibility that data would leak into the validation set because only the training set is over sampled. Oversampling balances the unbalanced dataset.

B. Reported Results

Based on the adopted architecture which is a customised Alex Net with Dense Layers, the results obtained have been reported in a tabular fashion.

Fig. 2 shows the light curve of an exoplanet whose ID is KIC 6922244.

Table I depicts the performance metrics in the form of Precision, Recall and F1-score that have been calculated during expolanet detection whereas Table II showcases the confusion matrix that summarizes the performance of the proposed classification model where instances in terms of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) are represented for the actual and predicted values of a planet being an exoplanet or a non-exoplanet.

Fig. 3 depicts the training and validation accuracy (Y-axis) obtained with the increasing number of epochs (X-axis) while Fig. 4 shows the training and validation loss (Y-axis) against the increasing number of epochs (X-axis).

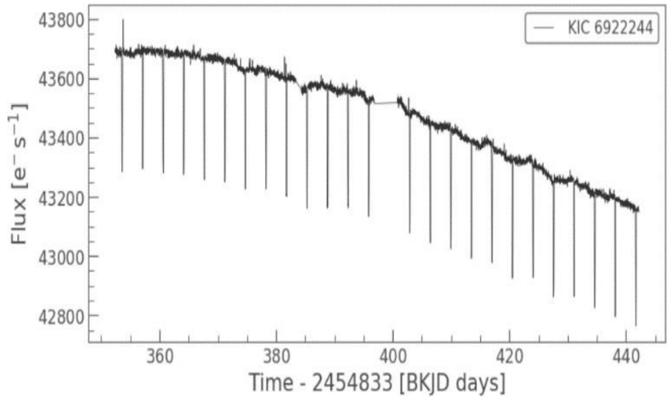

Fig. 2 Light curve of an exoplanet having ID– KIC 6922244

TABLE I
Performance Metrics

Label	Precision	Recall	F1-Score
0 (Non-Exoplanet)	1.00	0.97	0.98
1 (Exoplanet)	0.21	1.00	0.34

TABLE II
Confusion Metrics

Label	0 (Non-Exoplanet)	1 (Exoplanet)
0 (Non-Exoplanet)	546 (TN)	19 (FP)
1 (Exoplanet)	0 (FN)	5 (TP)

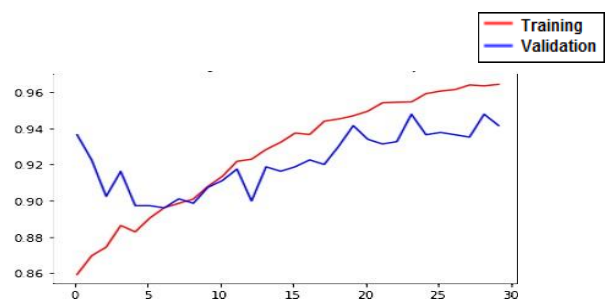

Fig. 3 Training and Validation Accuracy

https://doi.org/10.15864/ajac.21023

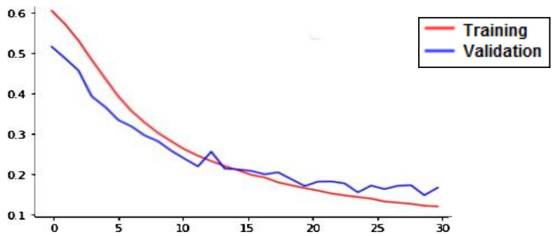

Fig. 4 Training and Validation Loss

C. Discussion & Analysis

It can be easily concluded that the proposed approach manages to correctly detect all the exoplanets without misclassification and 546 out of 565 non-exoplanets are also correctly recognized. In light of this, we can infer that while the proposed model does a good job of recognizing exoplanets, it makes trivial mistakes when categorizing non-exoplanets. The performance of the proposed exoplanet detection classifier can be evaluated from the confusion matrix parameters in terms of the accuracy score using (1):

$$Accuracy = (TP+TN) / (TP+TN+FP+FN) \qquad (1)$$

where *TP* stands for True Positives; which is the number of samples correctly identified as positive by the model, *TN* stands for True Negatives; which is the number of samples correctly identified as negative by the model, *FP* stands for False Positives; which is the number of samples incorrectly identified as positive by the model and *FN* stands for False Negatives; which is the number of samples incorrectly identified as negative by the model.

Thus, the model is performing quite well with an accuracy score of 96.67%.

V. Conclusion

The proposed work aims to identify, detect and classify exoplanets and non-exoplanets in the solar system using the Transit Photometry method by building a Convolutional Neural Network (CNN) model. The classification accuracy is found out to be 96.67%, which is a good indicator of the research endeavor being in the right direction. Thus, with the confluence of technology and Machine Learning tools, we can surely look forward to myriad explorations, thereby unveiling what lies beyond in the metagalactic space.

VI. Future Work

The technology of exoplanet transit detection is crucial for detecting new planets in the field of applied astrophysics [10]. In the search for exoplanets using Machine Learning techniques, one of the key factors that can significantly affect the performance and accuracy of the models is the availability of clean and high-quality datasets. Testing our proposed model on fresh, larger datasets would be worthwhile. Remember that the light curves of thousands of additional stars that the Kepler's telescope has gathered can be found in the MAST database (Mikulski Archive for Space Telescopes). But this data must be downloaded and processed in order to extract the light curves because they are not currently in a format that a Machine Learning model can use directly. Due of time constraints, it was not possible to test the proposed models using fresh datasets. Radio backscatter techniques have clearly shown signatures of O+. However, radio study still requires a lot of development. In addition to the availability of clean data, the use of more efficient algorithms is also critical in improving the performance of machine learning models for exoplanet detection. The XGBoost algorithm, for instance, is a popular and powerful technique that can be used to boost the performance of a classification model by combining the predictions of multiple weak classifiers. By using XGBoost techniques, the probability of achieving better outcomes is high, even when a relatively small set of classifiers is used.

VII. References

[1] A. Chaushev, L. Raynard, M.R. Goad, P. Eigmüller, D.J. Armstrong, J.T. Briegal, M.R. Burleigh, S.L. Casewell, S. Gill, J.S. Jenkins, L.D. Nielsen, C.A. Watson, R.G. West, P.J. Wheatley, S. Udry and J.I. Vines. (2019). Classifying exoplanet candidates with convolutional neural networks: application to the Next Generation Transit Survey. *Monthly Notices of the Royal Astronomical Society*. *488(4)*, pp. 5232-5250. Available: https://doi.org/10.1093/mnras/stz2058

[2] G. Olmschenk, S.I. Silva, G. Rau, R.K. Barry, E. Kruse, L. Cacciapuoti, V. Kostov, B.P. Powell, E. Wyrwas, J.D. Schnittman and T. Barclay. (2021). Identifying Planetary Transit Candidates in TESS Full-frame Image Light Curves via Convolutional Neural Networks. *The Astronomical Journal. 161(6)*, pp. 273. Available: https://iopscience.iop.org/article/10.3847/1538-3881/abf4c6

[3] L.C. Yeh and G. Jiang, (2020). Searching for Possible Exoplanet Transits from BRITE Data through a Machine Learning Technique. *Publications of the Astronomical Society of the Pacific*. *133(1019)*, pp. 014401. Available: https://doi.org/10.1088/1538-3873/abbb24

[4] D. Mislis, S. Pyrzas and K.A. Alsubai. (2018). TSARDI: a Machine Learning data rejection algorithm for transiting exoplanet light curves. *Monthly Notices of the Royal Astronomical Society*. *481(2)*, pp. 1624-1630. Available: https://doi.org/10.1093/mnras/sty2361

[5] I. Priyadarshini and V. Puri. (2021). A convolutional neural network (CNN) based ensemble model for exoplanet detection. *Earth Science Informatics. 14*, pp. 735-747. Available: https://doi.org/10.1007/s12145-021-00579-5

[6] S. Seager. (2008). Exoplanet transit spectroscopy and photometry. *Space Science Reviews. 135*, pp. 345-354. Available:https://doi.org/10.1007/s11214-008-9308-5

[7] A.C. Cameron, D.M. Wilson, R.G. West, L. Hebb, X.B. Wang, S. Aigrain, F. Bouchy, D.J. Christian, W.I. Clarkson, B. Enoch, M. Esposito, E. Guenther, C.A. Haswell, G. Hébrard, C. Hellier, K. Horne, J. Irwin, S.R. Kane, B. Loeillet, T.A. Lister, P. Maxted, M. Mayor, C. Moutou, N. Parley, D. Pollacco, F. Pont, D. Queloz, R. Ryans, I. Skillen, R.A. Street, S. Udry and P.J. Wheatley. (2007). Efficient identification of exoplanetary transit candidates from SuperWASP light curves. *Monthly Notices of the Royal Astronomical Society. 380(3)*, pp. 1230-1244. Available: https://doi.org/10.1111/j.1365-2966.2007.12195.x

[8] A. Krizhevsky, I. Sutskever and G.E. Hinton. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM. 60(6)*, pp. 84-90. Available: https://doi.org/10.1145/3065386

[9] M. Mendillo, P. Withers and P.A. Dalba. (2018). Atomic oxygen ions as ionospheric biomarkers on exoplanets. *Nature Astronomy. 2(4),pp.* 287-291. Available: https://doi.org/10.1038/s41550-017-0375-y

[10] A. Panahi, S. Zucker, G. Clementini, M. Audard, A. Binnenfeld, F. Cusano, D.W. Evans, R. Gomel, B. Holl, I. Ilyin, G.J. de Fombelle, T. Mazeh, N. Mowlavi, K. Niwnartowicz, S.S. Rimoldini and L. Eyer. (2022). The detection of transiting exoplanets by Gaia. *Astronomy & Astrophysics. 663*, pp. A101. Available: https://doi.org/10.1051/0004-6361/202243497

[11] M. Jara-Maldonado, V. Alarcon-Aquino, R. Rosas-Romero, O. Starostenko and J.M. Ramirez-Cortes. (2020). Transiting exoplanet discovery using machine learning techniques: a survey. *Earth Science Informatics. 13*, pp. 573-600. Available: https://doi.org/10.1007/s12145-020-00464-7

[12] Y. Wang, "The Identification of Transiting Exoplanet Candidates based on Convolutional Neural Network," In *Proceedings of the 2020 2nd International Conference on Big Data and Artificial Intelligence*, 2020, pp. 5-8, https://doi.org/10.1145/3436286.3436288

[13] Kaggle.com (2016). Exoplanet Hunting in Deep Space: Kepler labelled time series data. Available: https://www.kaggle.com/datasets/keplersmachines/kepler-labelled-time-series-data. Last accessed on 15/03/2023.

VIII. BIOGRAPHIES

Gaurav is currently pursuing Bachelor of Engineering in the Department of Computer Science & Engineering, University Institute of Technology, The University of Burdwan, West Bengal, India. His main research work focuses on Image Processing, Machine Learning and Deep Learning.

Sumit Gupta is currently working as Assistant Professor in the Department of Computer Science & Engineering, University Institute of Technology, The University of Burdwan, West Bengal, India. He has an Academic Experience of 11 years and a Research Experience of more than 8 years. He has published more than 30 articles, book chapters, research papers in reputed international and national conferences including IEEE, Springer, IETE, CRC Press etc. His main research work focuses on Natural Language Processing, Machine Learning, Deep Learning and Information Security.