Beyond the Stars: A Data-Driven Approach to Exoplanet Categorization

Prithika Chauhan

Abstract

Exoplanet detection improves the current understanding of planet formation and provides the possibility for the discovery of new habitable worlds, However, teams of astronomers and astrophysicists have traditionally been the only ones capable of identifying exoplanets. Using traditional techniques such as the Transit Method, Gravitational Microlensing, Direct Imaging Polarimetry, Astrometry, and Radial Velocity, researchers have attempted to identify exoplanets in the past, but manual processing is difficult and time-consuming. An advanced approach to detecting exoplanets in space is by utilizing artificial intelligence to solve this problem. In this study, hybrid approaches with an optimized elastic-net-based driven model were developed to detect exoplanets effectively. The standard planet dataset was collected and pre-processed to eliminate unnecessary components that can affect the prediction accuracy. The quality of the dataset was enhanced by the suggested framework's use of two distinct pre-processing techniques, including mass imputation and median and median absolute deviation-based normalization. To forecast the exoplanet, the pre-data was further analyzed using a feature selection procedure. The optimized elastic net was used to carry out the feature selection. These features were then further processed for prediction using a hybrid Bidirectional Gated Recurrent Units and Support Vector Machine (BiGRU-SVM) classifier. The performance metrics, such as accuracy, error, TNR, and F1-score, are evaluated in the evaluation of the proposed model.

Introduction

The quest to understand exoplanets, or planets outside our solar system, has become one of astronomy's most intriguing and active fields. The discovery of exoplanets has accelerated with the aid of advanced telescopes and observation techniques. As the number of candidate exoplanets increases, there is a growing need to analyze these datasets collected from advanced telescopes. Telescopes churn out millions of data points that are not feasible to understand through manual work done by scientists. Teams of scientists within the field need a more efficient way to understand these datasets. One extremely effective way to do this is by employing machine learning algorithms, which can comb through the hundreds of thousands of data points collected by telescopes and make predictions by identifying patterns in the data.

Machine learning, a subfield of artificial intelligence, has made significant strides in astronomy. It offers the potential to automate and enhance the identification of exoplanets based on available data. They analyze data and identify patterns. Then, based on these patterns, models make predictions and test those predictions on unseen data sets, making them more efficient than manual work.

Through the combination of these two fields, machine learning and exoplanet detection, scientists are just beginning to learn more about this intersection (Horner et al.). Although extremely useful in this field, machine learning has yet to be fully developed as a common technique to employ in astronomy. Therefore, the relationship between machine learning and exoplanet detection research has significant gaps.

Literature Review

Before fully comprehending the applications of machine learning on exoplanet detection, I needed to understand how machine learning works at a basic level. This meant understanding how they are formed, how they look for patterns in data, and what models are commonly used. Additionally, I had to understand the shortcomings of each model and how to develop my model for this field of analysis. Also, understanding how data is observed from exoplanets was vital to the initial research.

Exoplanet Background

To understand the use of machine learning when looking at exoplanets, it is important to understand what exoplanet data is composed of.

Methods of Discovery

There are many methods of discovery for finding exoplanets in deep space. One of the pioneering techniques in exoplanet discovery is the transit method. Gavin Ramsay and others, scientists at the Armagh Observatory and Planetarium, believe that some exoplanets reveal themselves through the timing variations they induce in the arrival of pulses from pulsars, rapidly rotating neutron stars (Ramsay et al.). The slight perturbations in the regular pulsar signals betray the gravitational influence of an unseen exoplanet. This method is what the Kepler space telescope used, according to NASA (NASA Exoplanet Archive). NASA's Kepler spacecraft spent over four years collecting this data on hundreds of thousands of star systems. Images were collected and the pixels corresponding to stars were identified, and intensity and location were identified as well over a set period. Putting all these together generated the light curves for each star from which an exoplanet could be detected. For example, when a planet passes in front of a star, the brightness of that star is observed and becomes dimmer. The data will show a dip in flux if a planet is transiting the star as explained by Mousavi-Sadret and others in their study Revisiting Mass-Radius Relationships for Exoplanet Populations: A Machine Learning Insight (Mousavi-Sadret al). In a real-world light curve, the data appears more mangled and with several systematic uncertainties that need to be subtracted, so ensuring that the data is acceptable for machine learning is vital, especially for building models.

Machine Learning Algorithm Basics: Learning Techniques

Machine learning is the process through which computers "learn" without being explicitly programmed as explained by Bhamare and others, who presented this information at the International Conference on Intelligent Technologies (Bhamare et al.). Complementary are typically used to rule out false positives and confirm that the signal identified as an exoplanet is not the result of a false positive source. On the other hand, Priyadarshini and Puri, writing in the *Earth Science Informatics* journal, state that advances in statistical and machine learning methods have led to their dependence on a novel procedure known as "validation," which was created to find new exoplanets (Priyadarshini and Puri). Rather than depending on fresh observations to enhance the transit method, the recently discovered exoplanets are verified through previously created machine learning methods that emulate the neural networks found in the human brain. Oltjon Kodheli and other scientists at the University of Luxembourg affirm that machine learning increases the accuracy score. Therefore we can have greater confidence when new signals are detected from stars already identified as exoplanets (Kodheli et al.).

This is extremely important for large data sets and making predictions or classifications for future data points. Within machine learning, four main learning techniques are commonly used. However, within exoplanet discovery, the techniques used are supervised learning and semi-supervised learning, as explained by Serjeant and others in *Nature Astronomy Vol. 4*. Supervised learning is using data to train the computer to infer a result. Essentially, it teaches the computer to make predictions based on the given data (Serjeant et al.). This technique is used so frequently because it allows for easier classification. By looking at already classified data, this technique allows the prediction to be made using the actual data through a model. The second most common technique is semi-supervised learning. This is a more realistic approach, due to the irregularities within the data. Many times a hybrid approach is necessary. Yucheng Jin and others, scientists at the University of California-Berkeley, explain that semi-supervised learning takes labeled and unlabeled data, meaning data that has already been classified and not classified (Jin et al.). It then takes this data to make a better prediction model. This technique is also helpful when labeling data, as seen many times with exoplanet classification. Overall, the fundamentals of machine learning within this field are built on these two techniques.

Machine Learning Algorithm Basics: Models

Models are the foundation of real-life applications within machine learning. In this field of study, the most used models are the logistic regression model, the K-Nearest Neighbors classifier, the decision tree model, and the random forest model, as highlighted by Manas Biswal in the Acceleron Aerospace Journal. These four models are all very similar, but their accuracy is vastly different. To start, the logistic regression model is one of the easiest to implement and train, however, it has the lowest precision score (Biswal). Precision within models is especially important because it ensures that the model is consistent and gives a correct representation of the data. Precision is calculated by dividing the number of true positives by the total number of predictions, according to Dr. Audenaert and a team of researchers at the Massachusetts Institute of Technology (Audenaert et al.). True positive refers to how many predicted values are classified as actual positive values. For example, if the model predicts that a value is positive, and the test data confirms that, that is a true positive. The higher the precision of the model, the higher the quality, which will allow for a more accurate prediction. Unfortunately, the logistic regression did not have a high precision score, meaning that the data was being overfitted. Overfitting refers to the model being unable to generalize the data. The logistic regression was unable to find patterns and instead just followed the training data.

The second model is the k-nearest neighbors classifier. This model had a similar precision and accuracy score. However, this model experienced the same issue of overfitting. The decision tree had a significantly higher accuracy score, however, the noise within the data set was affecting the model, causing the accuracy and precision scores to be vastly different each time it was run. Finally, the Random Forest model had the highest stability (Biswal). All in all, these models all have one purpose, to make predictions based on training data. However, each model comes with its problems, and due to the complexity of data sets seen in the astronomy field, these problems can be detrimental to the model.

Data Preprocessing Techniques for Machine Learning Algorithms

Data preprocessing is vital to any data set, especially when it comes to those related to astronomy. First, one of the most common problems that data preprocessing solves is an unbalanced data set. Within a data set, there is a majority and a minority class. This is a huge problem as it will skew the data, causing an imbalance within the classification data set. However, according to researchers at the Cochin University of Science and Technology, one way to remedy this problem is by using an over-sampling method called Synthetic Minority Over-Sampling Technique or SMOTE. This creates synthetic data to equal the minority class to the majority class (Agnes et al.). This is extremely helpful as it balances out the data set, allowing for the model to find true patterns, which will lead to a higher precision score. Digvijay Patil and others present an exoplanet identification technique based on machine learning's Random Forest Classification model in the *International Research Journal of Engineering and Technology*. When combined with the SMOTE preprocessing stage, the Random Forest Classifier Model was unable to predict values correctly, with less than 50% accuracy whether a light fluctuation had exoplanets (Patil et al.). Moreover, this method of preprocessing is not accurate and causes the accuracy to be off due to it overcompensating the minority class. So, even though the Random Forest Model performed well for comparative analysis against other models (Biswal) when isolated with a vital preprocessing technique, it severely underperformed according to Patil et al.'s study.

Another vital preprocessing technique is mass imputation, which is the process of imputed values development for categories through information integration. This is the approach for dealing with missing data from the response survey in the sample according to Abhishek Malik in the *Monthly Notices of the Royal Astronomical*

Society Journal. The primary goal is to create a single synthetic dataset of proxy values for the unobserved data in Sample A and apply it to the associated patterns of Sample B to produce projection estimators of population mean (Malik et al.). This is particularly useful when Sample A is a large-scale survey and data entry is very expansive to measure, such as exoplanet datasets. The proxy values are generated by fitting a working model related to the data from Sample B. Then, the synthetic values are generators, and thus Sample B is used as a training sample for predicting Sample A. Finally, the large blocks of missing values in a dataset are generated. This preprocessing technique was used extensively in the study done by Christopher Fluke and others, researchers at Swinburne University of Technology. The results of this study were extremely effective; by using a logistic regression model they achieved a 95% accuracy score and a precision score of 86% (Fluke et al.). Therefore, a logistic regression model greatly improves through the usage of this technique. Performing without this, the regression performs at a much lower value, as seen in the study with Biswal. Pratyush and Gangrade, in the journal article "Automation of Transiting Exoplanet Detection, Identification and Habitability Assessment Using Machine Learning Approaches", also experienced this with their Random Forest Model. With the use of this technique, the effectiveness of the model increased significantly, going from 50% accuracy to 92% accuracy (Pratyush and Gangrade). This demonstrates that preprocessing techniques can greatly improve the results of the models, but also that the variance between each trial of the model is vastly different.

Another technique used is the Mean and Standard Deviation-based Normalization Method techniques. The raw data's statistical mean and/or standard deviation are used to normalize the data. Zahra Ahmed and others from Standford University provide several variations to rescale or modify the data using these metrics. It consists of Z score normalization, which is the mean and standard deviation measures used to rescale the data such that resultant features have zero mean and unit variance (Ahmed et al.). In each instance, X of the data is transformed into Z score to find the mean and standard deviation of the feature. During this technique, data offset for each feature is calculated and the scaling factor is computed using an algorithm that gives information about the variation of more recent features.

While reducing the level of noise in the data, the approach enhances the representation of less concentrated characteristics. Additionally, it preserves some of the data's structure while eliminating the unit variance restriction, a feature that the Z score technique does not, according to Anne Dattilo in *The Astronomical Journal*. The square root of the raw data is computed, and the mean cantered approach is then used to rescale the data. The data is moved such that the highest value equals the difference between the feature's two extreme values and the minimum value coincides with zero (Dattilo). This process's primary drawback is that it is unable to produce additive multiplicative effects on the data. The multiplicative effect occurs when data has a standard deviation proportional to its mean. Abdul Karim and a team of researchers at the Noakhali Science and Technology University believe that except for the Z score method, these techniques assist in lessening the impact of outliers in the data but do not completely solve the issue of prominent features (Karim et al). But both mean and standard deviation measurements can change over time; none of the aforementioned techniques scale or convert data into an even numerical range. Also, this technique has yet to be used against an exoplanet dataset, so there is no testing of its effectiveness.

Overall, these two problems are most commonly seen when dealing with models, specifically with classification data sets, and preprocessing leading to misinterpretation and a lack of a standard in the industry. Currently, the focus regarding exoplanets is classification, so there is a clear potential for misinterpretation when it comes to analyzing these models and their predictions. There needs to be a standard created for these two aspects of machine learning to ensure relatability for these vital models.

The Connection

Machine learning algorithms excel at feature extraction and pattern recognition, enabling them to discern complex relationships within exoplanet datasets. As described by Jeroen Airapetian and others in The Astronomical Journal, in exoplanet classification, these algorithms analyze light curves, which represent the brightness of stars over time (Airapetian et al.). In The European Physical Journal: Special Topics, Margarita Safonova states that by learning patterns associated with confirmed exoplanets, machine learning models can then identify potential candidates in new datasets, streamlining the identification process (Safonova). The trained model can then autonomously classify new, unlabeled light curves based on the patterns it has learned. The transit method, a primary technique for exoplanet discovery, involves detecting the periodic dimming of a star's light caused by a planet passing in front of it. Machine learning algorithms can enhance transit detection by differentiating between genuine exoplanet signals and false positives, such as instrumental noise in the star's brightness. This reduces the likelihood of misclassifying non-exoplanetary phenomena as planets, emphasizing the importance of automating this analysis.

Gap in Research

Deep space observations often yield data that is both noisy and sparse. Machine learning algorithms traditionally perform better with clean, well-structured datasets. However, the inherent challenges of deep space observations, such as low signal-to-noise ratios and irregular data patterns, necessitate the development of specialized algorithms capable of robust performance in the face of such challenges. Machine learning algorithms, with their capacity to handle large datasets, analyze intricate patterns, and detect anomalies, represent a crucial bridge in overcoming this gap. However, there still needs to be more clarity about which model to use and how to work around its negatives to produce a quality model with a high understanding of the data. This leads to the questions, what algorithmic approaches can be developed to effectively classify exoplanets based on the analysis of light fluctuations from their host stars, and how do these approaches compare to existing classification methods?

Dataset

The dataset used in this study originates from the Kepler Space Telescope and was accessed through Kaggle.com. It records changes in the light intensity, or flux, of thousands of stars, which can provide insights into the presence of exoplanets. Each star in the dataset is assigned a binary designation: "2" indicates that at least one exoplanet has been confirmed to orbit the star, while "1" signifies that no exoplanets have been detected. These classifications are determined by analyzing characteristic patterns in the light fluctuations. For instance, Figure 1 illustrates two examples of light curves: the graph on the left displays periodic dips in light intensity, corresponding to an exoplanet orbiting the host star, while the graph on the right shows no such pattern, indicating the absence of an exoplanet.

After downloading the dataset, it was divided into two subsets: one for training the machine learning models and the other for testing their performance. This division is critical to ensure that the models are evaluated on unseen data, preventing overfitting and assessing their ability to generalize to new scenarios. The training data was used to teach the models to recognize patterns in light curves, while the testing data was reserved to measure how well the models performed on data they had not previously encountered.

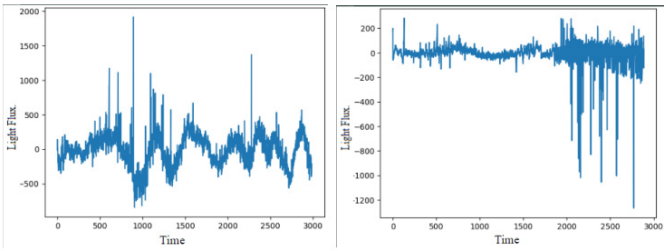

Fig. 1. Example light fluxes (in lumens) from the data set (Kaggle)

The dataset comprises over 3,000 data points, a volume that presents significant challenges for manual analysis. This difficulty becomes even more pronounced with larger datasets commonly used in exoplanetary research. The vast amount of information cannot be effectively examined by hand, making machine learning an indispensable tool. By automating the analysis, machine learning allows researchers to efficiently process these extensive datasets, uncovering patterns and making discoveries that would be infeasible through traditional methods. This dataset forms the foundation for evaluating the proposed binary classification model and conducting a comparative study of machine learning techniques in detecting exoplanets.

Methodology

In this research, there was a binary classifier model built which separated each light flux value into classes "Exoplanet" and "Non-Exoplanet". This exemplifies the purpose of the study, as it attempts to fill in the lack of a standard for which model is best and also have consistent results. As opposed to machine learning techniques applied for planet detection where features were determined automatically, features were derived from transformations and provided as inputs to an array of models for a comparative study. Based on the research done by other scholars in the field, the main cause of issues within the past used models came from the dataset leading to unwanted patterns and lack of reliability of which model is best. So, to combat this, the new binary classifier model was comprised of a combination of mass imputation and mean and standard deviation-based normalization methods, as well as a further developed model configuration. This meant being more complex, by running the data through the model multiple times as an already automated process. The model would repeatedly predict values and would use those proxy results to predict the values following. Figure 2 explains this process, showing the flow of each data point through the model. The goal of this process is to allow for a more accurate result.

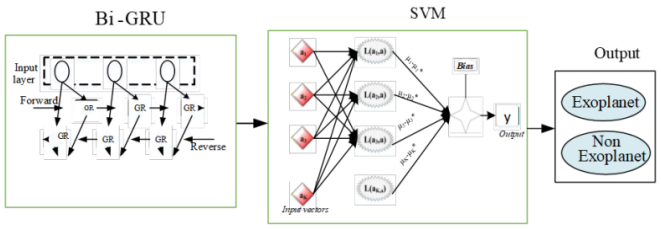

Fig. 2. Architecture of the new and proposed model

To test the effectiveness of the new model, an experimental design modeled after a comparative analysis was made. First, five models were selected to test the new model. As these five were the ones most commonly used, they were deemed best fit to test the new model's applicability. The chosen models include Long Short-term Memory-K-Nearest Neighbor (LSTM-KNN), Gated Recurrent Unit-Support Vector Machine (GRU-SVM), Recurrent Neural Network-Random Forest (RNN-RF), Convolutional Neural Network-Support Vector Machine (CNN-SVM), and Deep Neural Network-Random Forest (DNN-RF). The reasoning behind each of the models can be seen in the table below.

Model	Reasoning
LSTM-KNN	• Well-suited for sequential data processing • Ideal for capturing patterns in light curves • Simple and able to handle non-linear relationships
GRU-SVM	• Best used for sequential data processing and feature extraction
RNN-RF	• Capable of processing sequential data and extracting relevant features. • Able to handle noisy data
CNN-SVM	• Suitable for analyzing the structural features present in light curve data. • Able to handle high-dimensional data and non-linear classification boundaries.
DNN-RF	• Hierarchical feature representation • Allows for complex relationships to be learned from the data. • Can withstand overfitting
BiGRU-SVM	• Added one more layer as compared to the GRU-SVM Model

In order to test how effective these models are with the Kepler dataset, seven metrics were chosen. These are accuracy, precision, error, F1 score, true positive rate, true negative rate, and finally, false negative rate. The definitions of each of these can be seen in the following table.

Metric	Measured Quantity
Accuracy	• Percentage of correctly classified instances
Precision (PPV)	• The ratio of true positive predictions to the total predicted positives
Error	• The overall misclassification rate
F1 Score	• mean of precision and recall • Provides a balance between the two metrics
True Positive Rate (TPR)	• The proportion of actual positives that are correctly identified
True Negative Rate (TNR)	• The proportion of actual negatives that are correctly identified
False Negative Rate (FNR)	• The proportion of actual positives that are incorrectly classified as negatives

These metrics provide a comprehensive assessment of each model's performance in detecting exoplanets, accounting for both correct and incorrect classifications across different classes. In conclusion, the chosen models offer a diverse range of approaches to exoplanet detection, leveraging various ML techniques to analyze light curve data. Through rigorous evaluation using the selected metrics, we aim to identify the most effective model for automated exoplanet detection and contribute to advancements in the field of exoplanetary science.

Results and Analysis

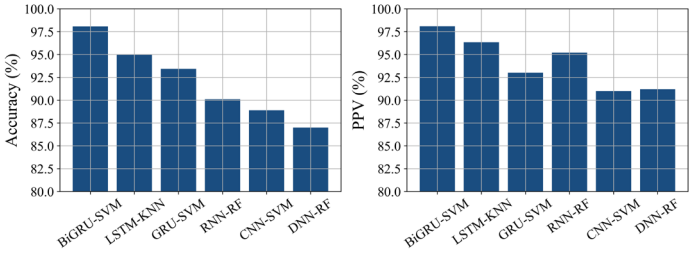

Fig. 3. Comparison of accuracy and PPV for models

In contrast to LSTM-KNN at 95.02%, GRU-SVM at 93.423%, RNN-RF at 90.1%, CNN-SVM at 88.9%, and DNN-RF at 87%, the Bi-GRU-SVM (my model) technique has an accuracy rate of 98.06%, showing that the new model is performing the best out of all the other selected models. The 5 other models are all performing at least 2% less accurately in comparison, showing that they are not as able to understand the data. This relationship is also seen with the precision or PPV value. No other model scores close to the new model, with the closest being the LSTM-KNN model and all others falling short.

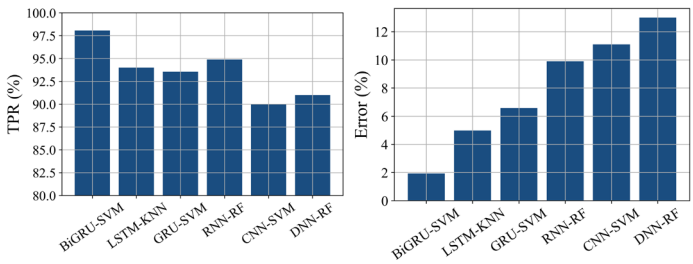

Fig. 4. Comparison of TPR and Error for models

Figure 4 presents a comparison of the percentage of positive cases that the model correctly classifies or TPR. The TPR of the BiGRU-SVM model is 98.05%, while that of LSTM-KNN is 94%, GRU-SVM is 93.56%, RNN-RF is 94.89%, CNN-SVM is 90%, and DNN-RF is 91%. These results are similar to the previous two metrics, where the new model is outperforming all the following models. Also, another trend can be seen, where the CNN-SVM is overall performing the worst out of the total six models used. This can be due to the total complexity of that particular model. Figure 4 also shows the error rates for different models. BiGRU-SVM model values are 1.9%, LSTM-KNNs are 4.98%, GRU-SVMs are 6.577%, RNN-RFs are 9.9%, CNN-SVMs are 11.1% and DN-RFs are 13%. Model performance decreases with a high error resulting in worse system performance and vice-versa. It also illustrates how the BiGRU-SVM model outperforms the traditional models in terms of error rates. Also, all the other metrics so far have almost opposite results, showing that the new model is, so far, still analyzing the data the most appropriately.

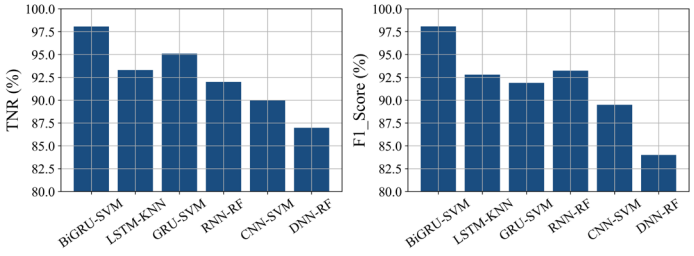

Fig. 5. Comparison of TNR and F1_score for Models

Figure 5 presents a comparison when the model accurately forecasts the negative class and F1 scores. The true negative rate (TNR), in comparison to LSTM-KNN at 93.3%, GRU-SVM at 95.1%, RNN-RF at 92%, CNN-SVM at 90%, and DNN-RF at 86.98%, the Bi-GRU-SVM technique has a TNR of 98.05%. Again, all the results support the hypothesis that the new model will perform the best in every chosen metric. After that, the F1 scores are examined for these models. The BiGRU-SVM technique has an F1 Score value of 98.07%, while LSTM-KNN is 92.8%, GRU-SVM is 91.89%, RNN-RF is 93.23%, CNN-SVM is 89.5%, and DNN-RF is 84%. So, the duality of the F1 score indicated that both the precision and accuracy of the new model are the most optimal, in conjunction with the others selected.

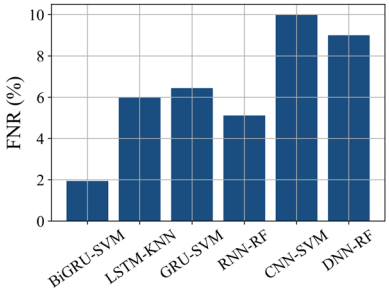

Fig. 6. Comparison of FNR models

The false negative rate (FNR) for models used in this study is shown in Figure 6. The FNR of the BiGRU-SVM model is 1.94%, while that of LSTM-KNN is 6%, GRU-SVM is 6.44%, RNN-RF is 5.11%, CNN-SVM is 10%, and DNN-RF is 9%. The FNR shows that the only

model that performs close to the new model is the RNN-RF, which has been underperforming in every other metric.

New Understanding

Overall, in every single metric, the BiGRU-SVM model functioned the best and had the highest overall efficiency. This model consistently beat the others, which emphasizes the fact that preprocessing had a significant effect on the results of the model. The other main issue with machine learning in astronomy is also solved, as it was a completely conclusive result, and there is no ambiguity as to which model is the most effective. A standard can now be set in the field because there is one model that works the best in every metric. The BiGRU-SVM model can provide the most holistic view of the data and provide the most accurate predictions without having the same issues that are seen with other models.

Limitations and Future Directions

The drawback of this method is that the angle between the planet's orbital plane and the direction of the observer's line of sight must be sufficiently small. Therefore, the chance of this phenomenon occurring is not high. More time and resources must be allocated to detect and confirm the existence of an exoplanet to ensure that there is enough data to back up the predictions made by the models used. Another limitation is the lack of fresh data. Given that the models were only tested on the singular dataset chosen, the applicability of the BiGRU-SVM model is limited. Therefore, if the research is to be used in the future, it is important to test the models on a wider variety of datasets. Another approach to this could be using a dataset that is composed of a different approach to exoplanet detection. Instead of light fluctuations being used, a good advancement of this could be images of the exoplanets instead. Overall, advancing this research is very important, and understanding the full capabilities of this new model can help us understand more about exoplanets in general.

Conclusion

In conclusion, the burgeoning field of exoplanet detection has witnessed a remarkable stride with the integration of machine learning techniques. Through this research, a comprehensive exploration of exoplanet detection methodologies using machine learning algorithms has been conducted, culminating in the development of a novel model exhibiting superior performance compared to five existing models across seven chosen metrics. The dataset, consisting of light flux data, served as the foundation for the research, enabling the training and testing of various machine learning models. The comparative analysis not only underscored the efficacy of machine learning in discerning exoplanetary signals amidst noise but also highlighted the significance of preprocessing techniques and model optimization in achieving accurate detection capabilities. The preprocessing techniques used for the datasets include mass imputation and normalization based on the median and median absolute deviation, the process of simultaneously completing a data file's significant missing block gaps.

At last, the hybrid Bi-GRU-SVM approach was applied to the classification of exoplanets. Furthermore, the effectiveness of the suggested method was examined and compared to other traditional methods, including LSTM-KNN, GRU-SVM, RNN-RF, CNN-SVM, and DNN-RF. With 98.06% accuracy, 1.9% error, 98.05% TNR, and 98.07% F1 score, we have proven that the suggested method performs well. It is evident from the comparison analysis that the proposed model yields a more efficient outcome than the current methods. Overall, the new model should be used as a new standard for preprocessing and model architecture for all future models used in the field due to its extremely effective ability to correctly detect exoplanets in deep space.

References

Agnes, C. K. , Naveed, A., and Chako, A. M. M. O. ExoSGAN and ExoACGAN: Exoplanet detection using adversarial training algorithms. *arXiv* (Cornell University), (2022, January). https://doi:10.48550/arxiv.2207.09665.

Ahmed, Z., D'Amico, S., Hu, R., & Damiano, M. (2021). Exoplanet detection from starshade images using convolutional neural networks. *Department of Aeronautics and Astronautics*, Stanford University, journal-article, slab.stanford.edu/sites/g/files/sbiybj25201/files/media/file/ahmed_spie2023_submitted.pd f.

Airapetian, V. S. , Barnes, R., Cohen, O. Collinson, G. A., Danchi, W. C., Dong, C. F., Del Genio, A. D., France, K., Garcia-Sage, K., Glocer, A., Gopalswammy, N., Grenfell, J. L., Gronoff, G., Gudel, M., Herbst, K., Henning, W. G., Jackman, C. H., Jin, M., Johnstone, C. P., ... Yamashiki, Y. (2019, July 2). Impact of space weather on climate and habitability of terrestrial-type exoplanets. *International Journal of Astrobiology*, vol. 19(2), 136-94. pp. 136–94, https://doi:10.1017/s1473550419000132.

Audenaert, J., Kuszlewicz, J. S., Handberg, R., Tkachenko, A., Armstrong, D. J., Hon, M., Kgoadi, R., Lund, M. N., Bell, K. J., Bugnet, L., Bowman, D. M., Johnston, C., Garcia, R. A., Stello, D., Molnar, L., Plachy, E., Buzasi, D., & Aerts C. (2021). TESS data for Asteroseismology (T'DA) stellar variability classification pipeline: setup and application to the kepler Q9 data. *The Astronomical Journal,* vol. 162 (5) 209. https://doi:10.3847/1538-3881/ac166a.

Bhamare, A. R., Baral, A., & Agarwal, Sl. (2021, June). Analysis of kepler objects of interest using machine learning for exoplanet identification." 2021 *International Conference on Intelligent Technologies (CONIT)*, https://doi:10.1109/conit51480.2021.9498407.

Dattilo, A. M., Vanderburg, A., Shallue, C. J., Mayo, A. W., Berlind, P., Bieryla, A., Calkins, M. L., Esquerdo, G. A., Everett, M. E., Howell, S. B., Latham, D. W., Scott, N. J., & Yu, L.. (2019, April) Identifying exoplanets with deep learning. Two New Super-Earths Uncovered by a neural network in K2 data. *The Astronomical Journal,* 157(5), 169. https://doi:10.3847/1538-3881/ab0e12.

Exoplanet hunting in deep space. (2017, April 12). *Kaggle,* www.kaggle.com/datasets/keplersmachines/kepler-labelled-time-series-data.

Fluke, C. J. & Jacobs, C. (2019, December 19). Surveying the reach and maturity of machine learning and artificial intelligence in astronomy. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery/Data Mining and Knowledge Discovery, 10 (2) https://doi:10.1002/widm.1349. Horner, J., Kane, S. R., Marchall, J. P., Dalba, P. A., Holt, T. R., Wood, J., Maynard-Casely, H E., Wittenmyer, R., Lykawka, P. S., Hill, M., Salmeron, R., Bailey, J., Lohne, T., Agnew, M., Carter, B. D., & Tylor, C. C. E. . (2020, September). Solar system physics for exoplanet research. *Publications of the Astronomical Society of the Pacific*, 132(1016) 102001.https://doi:10.1088/1538-3873/ab8eb9.

Jin, Y., Yang, L., and Chiang, C. (2022, April). Identifying exoplanets with machine learning methods: a preliminary study. *International Journal on Cybernetics and Informatics* 11(2), 31-42. https://doi:10.5121/ijci.2022.110203.

Karim, A., Uddin, J., & Riyad, M. M. H. (2021, January). Identifying important features

for exoplanet detection: A Machine Learning Approach." gphjournal.org, Jan. 2024, https://doi:10.5281/zenodo.10566250.

Kodheli, O., Lagunas, E., Maturo, N., Sharma, S. K., Shankar, B., Montoya, J. F. M., Duncan, J. C. M., Spano, D., Chatzinotas, S., Kisseleff, S., Querol, J., Lei, L., Vu, T. X., & Goussetis, G. l. (2020, October 1). Satellite communications in the new space era: A Survey and Future Challenges. *IEEE Communications Surveys and Tutorials/*, 23(1), 70-109. https:///doi:10.1109/comst.2020.3028247.

Manas Biswal, M. (2023, December) A short review on machine learning in space science and exploration. *Acceleron Aerospace Journal,* 1 (4) 84-87. https://doi:10.61359/11.2106-2317.

Ma, Y. & Wen, X. (2023, January). Shared rules between planetary orbits displayed by multi-planet systems and their function. Research Square (Research Square), https://doi:10.21203/rs.3.rs-2442884/v1.

Malik, A., Moster, B. P., & Obermeier, C. (2021, December). Exoplanet detection using machine learning. *Monthly Notices of the Royal Astronomical Society,* https://doi:10.1093/mnras/stab3692.

Mousavi-Sadr, M., Jassur, D. M., & Gozaliasl, G. (2023, August). Revisiting mass–radius relationships for exoplanet populations: a machine learning insight. *Monthly Notices of the Royal Astronomical Society*, 525 (3) 3469-85. https://doi:10.1093/mnras/stad2506. NASA Exoplanet Archive. exoplanetarchive.ipac.caltech.edu.

Digvijay, P., Srinivasarao, S. R., & Puri, R. (2021, June). Predicting the presence of exoplanets in star- systems using random forest classifier. *IJERT*, https://doi:10.17577/IJERTCONV9IS07026.

Pratyush, Pawel, and Akshata Gangrade. (2021, January). Automation of transiting exoplanet detection, identification and habitability assessment using machine learning approaches. *arXiv* (Cornell University), https://doi:10.48550/arxiv.2112.03298.

Priyadarshini, I. & Puri, V. (2021, February). A convolutional neural network (CNN) based ensemble model for exoplanet detection. *Earth Science Informatics,*14(2), 735-47. https://doi:10.1007/s12145-021-00579-5.

Ramsay, G., Kolotkov, D., Doyle, J. G., & Doyle, L. (2021, November) Transiting exoplanet survey satellite (TESS) observations of flares and quasi-periodic pulsations from low-mass stars and potential impact on exoplanets. *Solar Physics*, 296 (11) https://doi:10.1007/s11207-021-01899-x.

Safonova, M., Mathur, A., Basak, S., Bora, K., & Agrawal, S. (2021, July). Quantifying the classification of exoplanets: in search for the right habitability metric. *The European Physical Journal. Special Topics,* 230(10), 2207-20. https://doi:10.1140/epjs/s11734-021-00211.

Serjeant, S., Elvis, M., & Tinetti, G. (2020, November). The future of astronomy with small satellites. *Nature Astronomy*, 4(11), 1031-38., https://doi:10.1038/s41550-020-1201-5.