Exoplanet Detection with Radial Velocity Data

Tirth Surti
Department of Computer Science
Stanford University
tsurti@stanford.edu

Abstract

The radial velocity method is historically the second most successful exoplanet detection method, having led to the discovery of over 900 exoplanets. With future radial velocity surveys expected to achieve a precision to detect near Earth-like planets, novel techniques to identify exoplanets will become increasingly important. We introduce several deep learning techniques including fully-connected neural networks and 1D CNN-LSTM models to identify exoplanets from measured and simulated radial velocity data. We argue that these models provide a promising baseline for future exoplanet detection as more radial velocity data is obtained. We further show that the CNN-LSTM model enables for an exoplanet identification method with little preprocessing.

1 Introduction

Since the confirmation of the first exoplanet beyond our solar system in 1992, exoplanet searches have become one of the principle tasks of institutions like NASA, with over 5,000 confirmed exoplanets discovered since then (1). Identifying and characterizing exoplanets enables us to move closer to answering questions regarding the existence of life beyond Earth. Furthermore, it allows us to compare the distribution and properties of planets in our own solar system to that of others and understand how solar systems form and evolve. We thus aim to determine the efficacy of using deep learning frameworks for exoplanet detection.

One of the most popular methods used to detect exoplanets is by measuring the radial velocity of the host star that one or more exoplanets orbit around. The radial velocity of a star encodes how much it "wobbles" due to its orbit around the planetary system's center of mass, resulting in periodic movement of the host star. One of the main challenges with detecting exoplanets with radial velocity data, particularly for Earth-sized planets, is the existence of stellar noise due to magnetic activity that produces signals comparable to that of the planet. This has led to false positives in exoplanet detection (13).

Deep learning is yet to be used for the detection of exoplanets from radial velocity data and could be harnessed to identify underlying trends that traditional methods cannot. We aim to explore the performance of a fully-connected neural network that takes as input time-series extracted features from the radial velocity curves and a 1D CNN-LSTM model that directly takes as input the sequential time-series radial velocity data. The output is whether the radial velocity data corresponds to an exoplanetary system.

2 Related Work

Current methods for detecting exoplanets through radial velocity data is primarily done through signal processing and statistical analysis. The Lomb-Scargle Periodogram, a variation on the Discrete Fourier Transform, can be used to identify peaks in the frequency domain with maximum-likelihood estimation (14)(10). Monte Carlo methods have also been used to find maximum-likelihood parameters that model the radial velocity data (3). In general, such methods can be inefficient due to the necessity of exploring vast parameter spaces.

While deep learning is yet to be used for identifying exoplanets using the radial velocity method, it has been successful in removing stellar activity noise in radial velocity data, which could potentially aid future exoplanet detection frameworks (5). We find their method to generate simulated radial velocity curves robust and considerate of various types of stellar signals. Machine learning techniques have also been used to identify exoplanets through the more popular transit method, which looks at drops in the brightness of a host star due to a crossing planet. Malik et al. utilized a time-series feature extractor *tsfresh* on the transit data and trained a gradient boosting model to identify exoplanets (11). NASA has also developed a more comprehensive tool *ExoMiner* that uses a deep convolutional neural network, taking input of not only summarizing features but also of the original time series data (15).

We find NASA's *ExoMiner* to be the current state of the art method for exoplanet detection because it parallels the current methods used to detect exoplanets by hand; *ExoMiner* directly considers unfolded and phase-folded fluxes rather than summarized data. We similarly aim to use the time series data directly as input.

3 Dataset

Another significant challenge in applying deep learning to radial velocity data is the lack of confirmed planets detected through radial velocities. Therefore, we have not only applied deep learning methods on real radial velocity data but also simulated data, for which a larger amount of data can be obtained.

3.1 Real Data

A total of 1070 radial velocity curves of confirmed exoplanetary systems were retrieved from NASA's Exoplanet Archive (1). Additionally, a total of 4270 radial velocity curves of stellar systems not confirmed to have exoplanets containing at least two measurements were retrieved by querying the Data Analysis and Center for Exoplanets' online database (4). The non-exoplanetary radial velocity curves were obtained by ensuring that the corresponding star's name was not found in the database of all host stars containing exoplanets.

3.2 Simulated Data

A total of 20,000 simulated radial velocity curves, taking into account stellar activity, were generated with random sampling. Using the software *SOAP 2.0*, 20,000 radial velocity curves resulting from just stellar activity were generated using uniform randomly sampled star configurations, including stellar rotation periods, effective surface temperatures, and the size and number of active regions (6)(9)(8).

Using *RadVel*, a radial velocity fitting package, 10,000 random exoplanetary radial velocity curves were generated using uniform randomly sampled orbital parameters, including period, time of conjunction, eccentricity, argument of periastron, and radial velocity semi-amplitude (7). These curves were then injected with the stellar activity radial velocity curves generated from *SOAP 2.0*, resulting in 10,000 exoplanetary and 10,000 non-exoplanetary radial velocity curves. To simulate current instrumental precision and other random stellar effects not accounted for by *SOAP 2.0*, a Gaussian error of 2 m/s was added to the radial velocities.

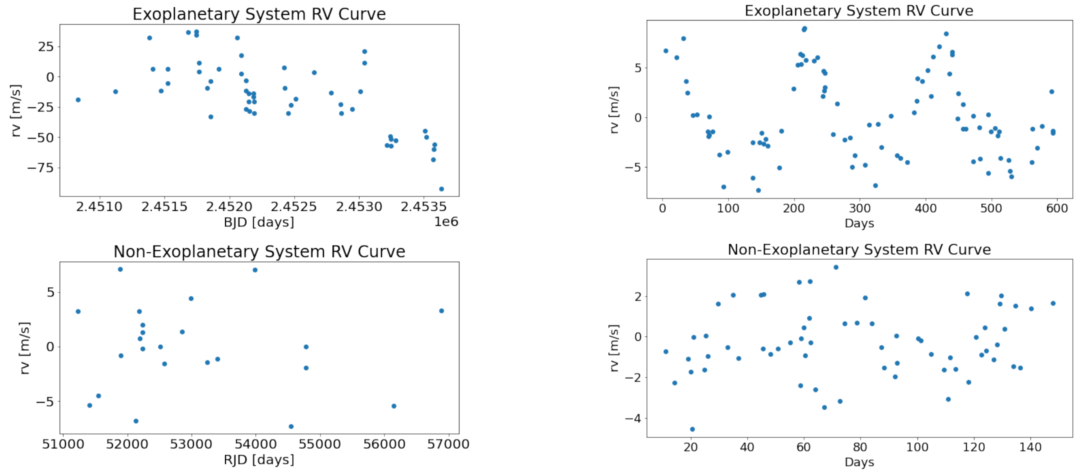

Figure 1: Sample radial velocity curves from the real dataset (left) and simulated dataset (right).

Given the small dataset sizes for a typical deep learning application, a 60-20-20 training, validation, and test split, respectively was used for both the real and simulated datasets.

4 Methods and Experiments

We use a standard neural network and 1D CNN-LSTM model for both the simulated and real radial velocity data. Since this is a binary classification task, we used the binary cross entropy loss function.

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\Big(y^{(i)}log(\hat{y}^{(i)}) + (1-y^{(i)})log(1-\hat{y}^{(i)})\Big)$$

For exoplanet detection, recall is the most important metric as we seek to recover as many exoplanet candidates as possible. Precision is stressed less as false positives might indicate the existence of unknown exoplanets in the data. Following that of Malik et al., since recall is dependent on the decision threshold, we instead first optimized hyperparameters of each model based upon the area under the ROC curve (AUC) so that we obtain a model that performs maximally over all decision thresholds (11). To optimize recall, we then chose the decision threshold value that maximized recall subject to the constraint that the F1 score, the harmonic mean of precision and recall, was at a minimum 0.7 so as to not neglect precision.

4.1 Fully-Connected Neural Network

For both the real and simulated radial velocity data, a time-series feature extractor *tsfresh* was used to obtain inputs as a fully-connected neural network cannot have as input raw radial velocity data. Sample features extracted include Fourier transform coefficients, partial autocorrelations, number of peaks, variances, and absolute maximums (12). This feature-extracted data was normalized feature-wise for both datasets.

For both the real and simulated datasets, hyperparameters to optimize AUC were chosen using Keras' tuner hyperband. Learning rates from the set $\{0.0001, 0.001, 0.01\}$ and for each hidden layer in a given model, unit sizes in the range (16, 1024) with uniform step size 16 were sampled for each data type and architecture.

4.1.1 Real Data FC Model

We found that a shallow neural network with a single hidden layer was sufficient enough to achieve low avoidable bias. We used L2 regularization on the hidden layer with $\lambda = 0.01$ to avoid overfitting. Batch gradient descent with the Adam optimizer was done, taking $lr = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Training was done with batch sizes of 32 over 15 epochs.

4.1.2 Simulated Data FC Model

Unlike that of the real data, we found that a deeper neural network with three hidden layers was sufficient enough to achieve low avoidable bias. Batch gradient descent with the Adam optimizer was done, taking $lr = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Training was done with batch sizes of 32 over 10 epochs.

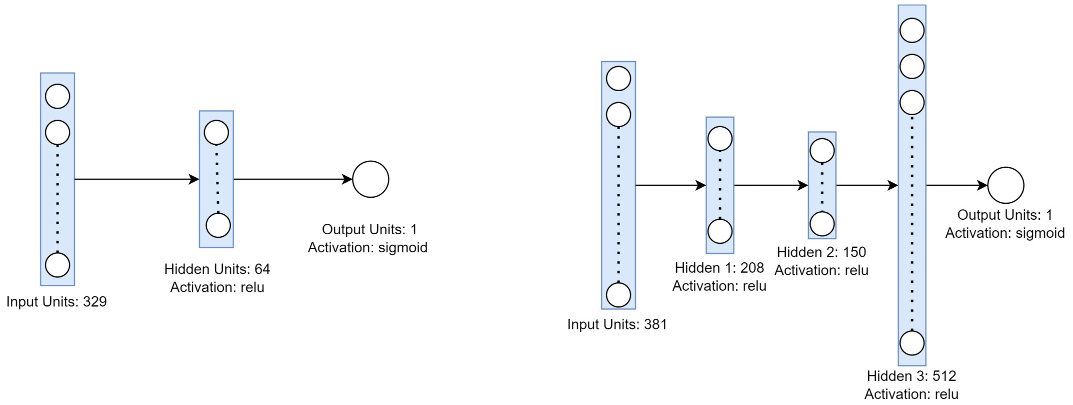

Figure 2: Fully-Connected Neural Network Architectures for Real Data (left) and Simulated Data (right)

Both fully-connected neural network models for the real and simulated data are summarized in Figure 2.

4.2 1D CNN-LSTM Model

One of the main disadvantages of feature-extracted inputs is that it requires significant computational resources to obtain relevant features, particularly with *tsfresh*. Furthermore, depending on the dataset, different number of relevant features may be extracted, requiring us to update the model architecture. A simpler pipeline involving the direct use of the radial velocity sequence data can be more efficient, which is enabled by a CNN-LSTM model.

For both the real and simulated datasets, the sequential radial velocity data was normalized and zero-padded or truncated to 100 datapoints. The CNN-LSTM models first pass this sequential data through convolution layers to extract features from the input sequences. These sequences are then passed into a many-to-one bidirectional LSTM layer as both previous and future datapoints provide information about the existence of an exoplanetary system. The LSTM activations are then passed through fully-connected layers for classification.

Similarly to optimize for AUC, the same learning rates and hidden unit sizes were sampled as the fully-connected neural network. However, due to random sampling inefficiencies for convolution and LSTM layers, the choice of convolution filter sizes and LSTM units were instead optimized through iteration based off of past utilizations of CNN-LSTM models for other time series classification problems that have worked well (2).

4.2.1 Real Data CNN-LSTM Model

Preliminary iteration showed that we needed two pairs of convolution layers before the bidirectional LSTM to achieve initial comparable performance to that of feature-extracted model. Batch gradient descent with the Adam optimizer was done, taking $lr = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Training was done with a batch size of 32 over 20 epochs. Figure 3 summarizes the model architecture for the real data.

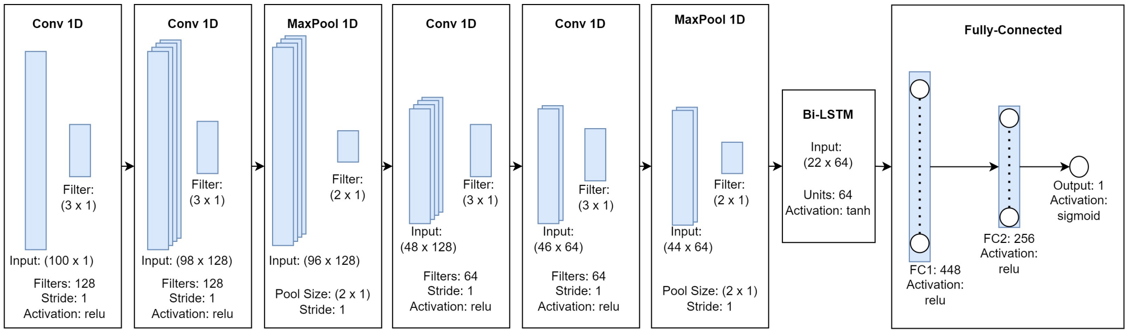

Figure 3: CNN-LSTM architecture for real data.

4.2.2 Simulated Data CNN-LSTM Model

Preliminary iteration for optimization showed that we needed just a single pair of convolution layers before the bidirectional LSTM to achieve comparable performance to that of the feature-extracted model. Batch gradient descent with the Adam optimizer was done, taking $lr = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Training was done with a batch size of 32 over 10 epochs. Figure 4 summaries the architecture for the simulated data.

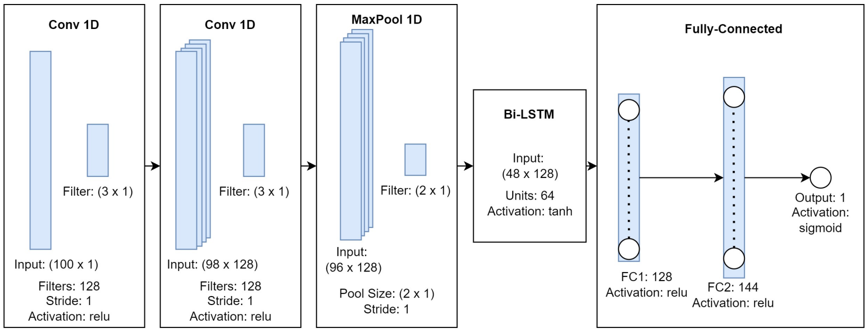

Figure 4: CNN-LSTM architecture for simulated data.

5 Results and Discussion

Decision thresholds for each dataset and architecture were chosen by analyzing the plots of recall, precision, and F1 score across various thresholds (see Figure 6). We note that for the real dataset and the CNN-LSTM model, an F1 threshold of 0.7 could not be reached, so the threshold was picked such that the F1 score was maximized. Table 1 summarizes the metrics across each model and Figure 5 shows the confusion matrix for each dataset and model type summarizing the precision and recall metrics.

	Model	AUC	Decision Threshold	Recall	Precision	F1 Score
Real Data	FC NN	0.938	0.200	0.869	0.599	0.709
	CNN-LSTM	0.8802	0.300	0.744	0.508	0.604
Simulated Data	FC NN	0.839	0.200	0.870	0.620	0.724
	CNN-LSTM	0.8601	0.260	0.911	0.597	0.721

Table 1: Summary of results across each model and dataset type.

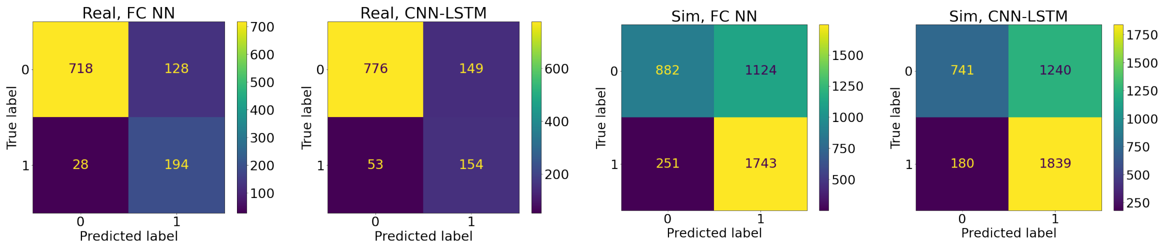

Figure 5: Confusion matrices for each dataset and model type, with '1' being an exoplanetary system and '0' representing a non-exoplanetary system.

For the real radial velocity data, the fully connected neural network achieved better performance than that of the CNN-LSTM model. Not only does the fully connected neural network have a larger capacity to discriminate between exoplanetary and non-exoplanetary systems due to the higher AUC, but it also achieved a higher recall for a significantly higher F1 score, suggesting the overall lack of performance of the CNN-LSTM model regardless of the threshold chosen. One potential reason for this difference likely comes from the input data quality. Many confirmed and non-confirmed exoplanets have significantly less than 100 datapoints, and applying convolutions on mostly zero-padded inputs may not result in as significant features as those produced by *tsfresh*.

However, for the simulated data, performance on the CNN-LSTM model is marginally better than the fully-connected neural network model, suggesting that the significance of features generated through the convolution and LSTM layers is comparable to that of *tsfresh*. It is also observed that in general, the simulated data did not achieve as high AUC as that of the real data. Reflecting what often occurs in practice, many simulated exoplanetary systems with smaller signals can be difficult to distinguish from non-exoplanetary systems with just stellar noise and are likely to result in similar features (see Figure 7). In practice, most confirmed exoplanetary systems have larger signals that are more easily distinguished.

6 Conclusion

Performance of the fully connected and CNN-LSTM models on relatively small datasets show significant promise in applying these techniques to identify exoplanets once more data becomes available in the future. We also see that the CNN-LSTM model, whose performance was comparable to that of the fully connected model, may have the potential to identify exoplanets with minimal preprocessing (no preliminary feature extraction needed) and supplement existing manual methods for identifying exoplanets from radial velocity data.

As it is likely that stellar noise is still a significant source of confusion for small signal exoplanetary systems, future work will likely involve connecting this exoplanetary identification pipeline with existing pipelines that process data to reduce stellar activity signals like that of de Beurs et al (5). This may enable use to recover a clearer radial velocity signal before applying further deep learning techniques for classification.

References

[1] NASA Exoplanet Archive. Planetary systems. 2022. URL: https://catcopy.ipac.caltech.edu/dois/doi.php?id=10.26133/NEA12, doi:10.26133/NEA12.

[2] Saksham Bassi, Kaushal Sharma, and Atharva Gomekar. Classification of variable stars light curves using long short term memory network. *Frontiers in Astronomy and Space Sciences*, 8, 2021. URL: https://www.frontiersin.org/article/10.3389/fspas.2021.718139, doi:10.3389/fspas.2021.718139.

[3] Floyd Bullard. *Exoplanet Detection: A Comparison of Three Statistics or How Long Should It Take to Find a Planet?* PhD thesis, Yale University Department of Statistical Science, 2009.

[4] The Data and Analysis Center for Exoplanets. Exoplanets Table, 2022. URL: https://dace.unige.ch/exoplanets/?filters={}.

[5] Zoe L. de Beurs, Andrew Vanderburg, Christopher J. Shallue, Xavier Dumusque, Andrew Collier Cameron, Lars A. Buchhave, Rosario Cosentino, Adriano Ghedina, Raphaëlle D. Haywood, Nicholas Langellier, David W. Latham, Mercedes López-Morales, Michel Mayor, Giusi Micela, Timothy W. Milbourne, Annelies Mortier, Emilio Molinari, Francesco Pepe, David F. Phillips, Matteo Pinamonti, Giampaolo Piotto, Ken Rice, Dimitar Sasselov, Alessandro Sozzetti, Stéphane Udry, and Christopher A. Watson. Identifying exoplanets with deep learning. iv. removing stellar activity signals from radial velocity measurements using neural networks, 2020. URL: https://arxiv.org/abs/2011.00003, doi:10.48550/ARXIV.2011.00003.

[6] X. Dumusque, I. Boisse, and N. C. Santos. SOAP 2.0: A TOOL TO ESTIMATE THE PHOTOMETRIC AND RADIAL VELOCITY VARIATIONS INDUCED BY STELLAR SPOTS AND PLAGES. *The Astrophysical Journal*, 796(2):132, nov 2014. URL: https://doi.org/10.1088%2F0004-637x%2F796%2F2%2F132, doi:10.1088/0004-637x/796/2/132.

[7] Benjamin J. Fulton, Erik A. Petigura, Sarah Blunt, and Evan Sinukoff. RadVel: The radial velocity modeling toolkit. *Publications of the Astronomical Society of the Pacific*, 130(986):044504, mar 2018. URL: https://doi.org/10.1088%2F1538-3873%2Faaaaa8, doi:10.1088/1538-3873/aaaaa8.

[8] G. M. H. J. Habets and J. R. W. Heintze. Empirical bolometric corrections for the main-sequence. , 46:193–237, November 1981.

[9] U. Heiter, P. Jofré, B. Gustafsson, A. J. Korn, C. Soubiran, and F. Thévenin. Gaia FGK benchmark stars: Effective temperatures and surface gravities. , 582:A49, October 2015. arXiv:1506.06095, doi:10.1051/0004-6361/201526319.

[10] Muhammad Salman Khan, James Stewart Jenkins, and Nestor Becerra Yoma. Discovering new worlds: a review of signal processing methods for detecting exoplanets from astronomical radial velocity data, 2016. URL: https://arxiv.org/abs/1611.00766, doi:10.48550/ARXIV.1611.00766.

[11] Abhishek Malik, Benjamin P Moster, and Christian Obermeier. Exoplanet detection using machine learning. *Monthly Notices of the Royal Astronomical Society*, dec 2021. URL: https://doi.org/10.1093%2Fmnras%2Fstab3692, doi:10.1093/mnras/stab3692.

[12] Julius Neuffer Andreas W. Kempa-Liehr Maximilian Christ, Nils Braun. tsfresh documentation, 2016-2021. URL: https://tsfresh.readthedocs.io/en/latest/index.html#.

[13] Lee J. Rosenthal, Benjamin J. Fulton, Lea A. Hirsch, Howard T. Isaacson, Andrew W. Howard, Cayla M. Dedrick, Ilya A. Sherstyuk, Sarah C. Blunt, Erik A. Petigura, Heather A. Knutson, Aida Behmard, Ashley Chontos, Justin R. Crepp, Ian J. M. Crossfield, Paul A. Dalba, Debra A. Fischer, Gregory W. Henry, Stephen R. Kane, Molly Kosiarek, Geoffrey W. Marcy, Ryan A. Rubenzahl, Lauren M. Weiss, and Jason T. Wright. The california legacy survey. i. a catalog of 178 planets from precision radial velocity monitoring of 719 nearby stars over three decades. *The Astrophysical Journal Supplement Series*, 255(1):8, jul 2021. URL: https://doi.org/10.3847%2F1538-4365%2Fabe23c, doi:10.3847/1538-4365/abe23c.

[14] J. D. Scargle. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. , 263:835–853, December 1982. doi:10.1086/160554.

[15] Hamed Valizadegan, Miguel J. S. Martinho, Laurent S. Wilkens, Jon M. Jenkins, Jeffrey C. Smith, Douglas A. Caldwell, Joseph D. Twicken, Pedro C. L. Gerum, Nikash Walia, Kaylie Hausknecht, Noa Y. Lubin, Stephen T. Bryson, and Nikunj C. Oza. ExoMiner: A highly accurate and explainable deep learning classifier that validates 301 new exoplanets. *The Astrophysical Journal*, 926(2):120, feb 2022. URL: https://doi.org/10.3847%2F1538-4357%2Fac4399, doi:10.3847/1538-4357/ac4399.

Appendix

A.1 Precision, Recall, F1 Score Tradeoff

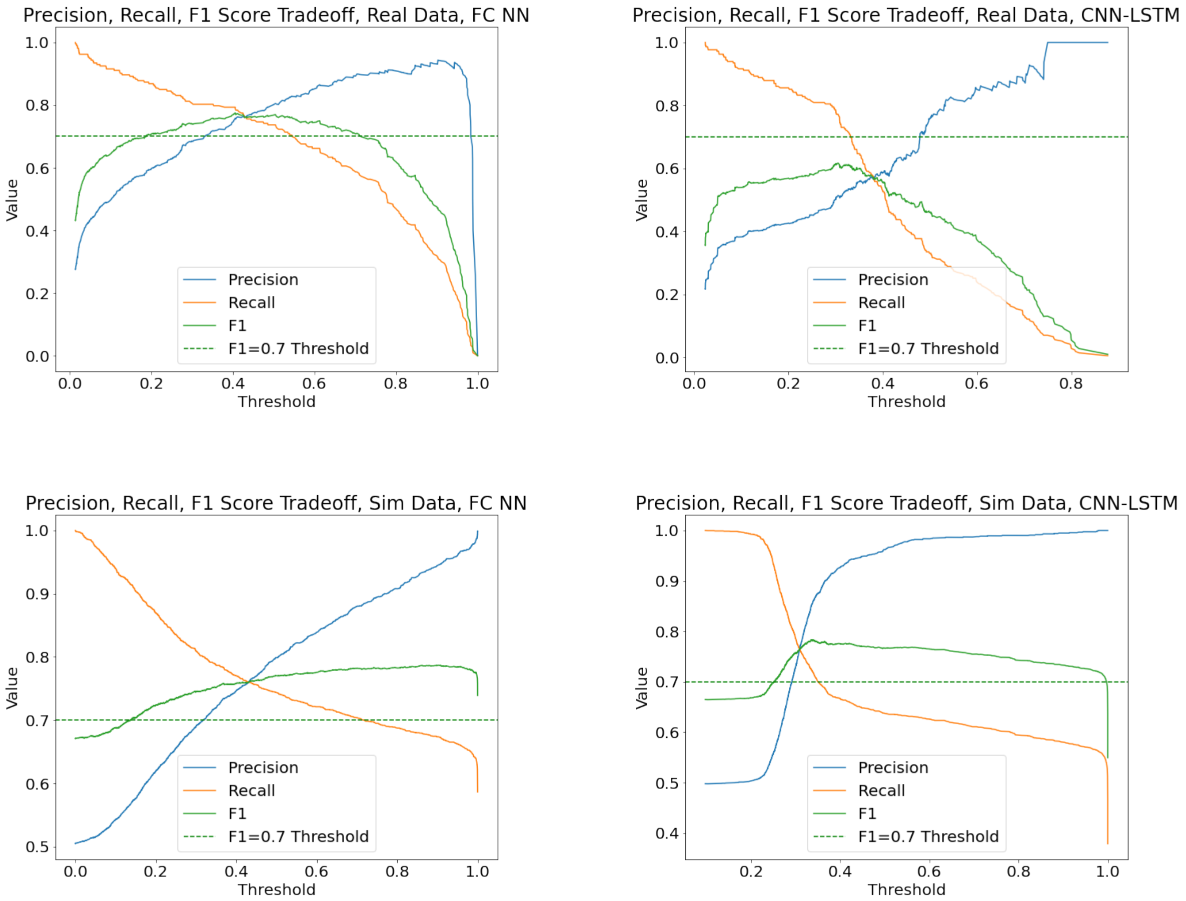

Figure 6: Precision, Recall, and F1 tradeoff curves for each dataset and model type. All of the models achieve a range of decision thresholds for which the F1 score is larger than 0.7 except for the CNN-LSTM model for the real data.

A.2 Small Signal Simulated Data

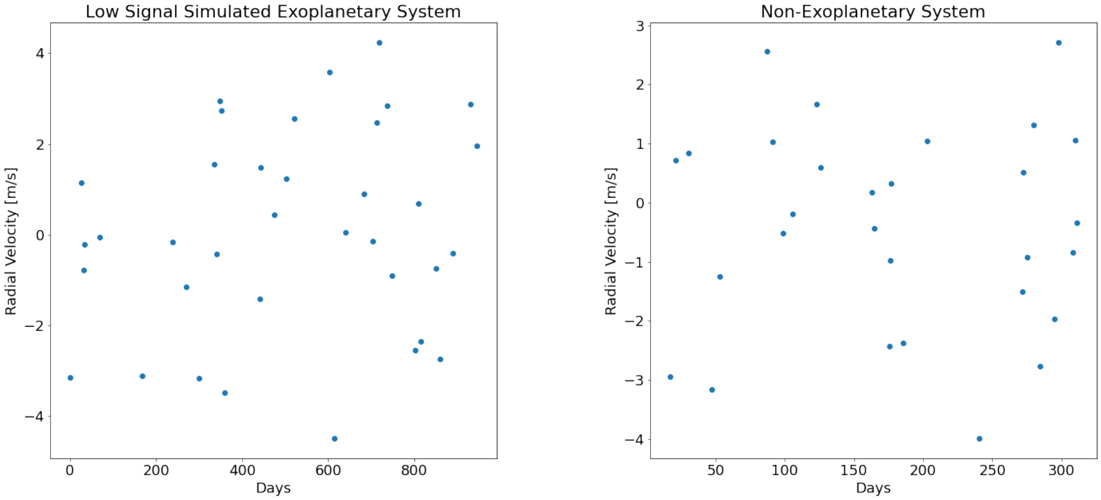

Figure 7: Example of a simulated exoplanetary system with low signal (left) compared to a non-exoplanetary system (right). For small signals, there's no longer clear periodicity in the data, making it very similar to the non-exoplanetary system's radial velocities.