**The Accuracy of Machine Learning Models in Stellar Brightness Prediction using GAIA DR2**

A colorful galaxy in space

Description automatically generated with medium confidence

Astronomical artwork generated by a machine learning model.

37256307

Word Count: 3,012

***Contents:***

[Abstract: 3](#_Toc171368489)

[Literature review: 3](#_Toc171368490)

[Scientific Research Question: 5](#_Toc171368491)

[Scientific Hypothesis: 5](#_Toc171368492)

[Methodology: 5](#_Toc171368493)

[Results: 6](#_Toc171368494)

[Discussion: 13](#_Toc171368495)

[*Data selection and ethical considerations:* 13](#_Toc171368496)

[*Choice of statistical testing and visualisation:* 13](#_Toc171368497)

[*Interpretation of results* 14](#_Toc171368498)

[*Limitations of the experiment and further implications of research:* 14](#_Toc171368499)

[Conclusion: 15](#_Toc171368500)

[References: 17](#_Toc171368501)

[Appendix: 20](#_Toc171368502)

# ***Abstract:***

The aim of this study is to assess the accuracy of different machine learning models in predicting the brightness of stars by training it on various stellar parameters provided by ESA’s GAIA dataset. Advancements in technology have affected the way astronomers study and view stars, from supercomputers to map stellar clusters from the ground to the new JWST able to observe various phenomena up to 13.6 billion light years away. By accurately predicting stellar parameters using machine learning, a new way of studying stars can be found. The study looks at 6 different machine learning models and assesses the accuracy of their predictions to an observed output. The methodology used judges the loss and correlation of the predicted output to an observed value, showing significant accuracy and minimal loss from each model.

# ***Literature review:***

Changeat et al.’s (2018) article provides a detailed analysis of different machine learning methods applied specifically to exoplanet atmospheric data. The study explores advanced machine learning techniques such as deep learning and ensemble methods, and different optimisation functions, however, the focus is primarily on atmospheric properties rather than direct stellar properties such as brightness. The study relies on specific sets of exoplanet data that were pre - determined, which may reduce the useability of the models in other astrophysical aspects. Published in the *Monthly Notices of the Royal Astronomical Society* (MNRAS), the article benefits from substantial academic rigor and credibility, making it a reliable source. However, it highlights a gap in applying machine learning prediction algorithms to specific stellar parameters. (Changeat, Tsiaras, Waldmann, & Tinetti, 2018).

Armstrong et al. (2019) address the prediction of stellar properties and stellar brightness fluctuations using machine learning regression models. The study utilizes a large stellar dataset of observations (unlinked), increasing the accuracy of predictions. However, the research focuses more on basic machine learning models with less detail on newer methods like deep learning and neural networks. The performance metrics used to evaluate the machine learning models are not discussed thus the model’s accuracy and loss are not able to be gauged, hence presenting a gap in the study. As a peer-reviewed article published in MNRAS, it is a reliable source that provides valuable insights into the application of machine learning for stellar brightness prediction (Armstrong, Kirk, Lam, McCormac, & Walker, 2019).

Jones et al. (2016) present a new and innovative approach to predicting stellar activity using machine learning. The study introduces new renditions of machine learning techniques used for astrophysics (ridge, polynomial and lasso regression), which can be compared with different stellar parameters such as brightness. Using streamed data from multiple stellar observatories (unlinked) enhances the scope and usability of the results and models produced. However, the complexity of the methods used may limit accessibility to non-academic and casual researchers. While the focus is on stellar activity, the research touches upon stellar parameters but does not go into any more detail. Published in the *Astrophysical Journal*, the article is a reliable and thorough study in both fields of astrophysics and machine learning (Jones, Smith, & Brown, 2016).

White et al. (2022) focus on variable star classification, a step prior to predicting stellar properties. The study evaluates multiple machine learning algorithms such as KNN, logistic regression and decision trees, providing a comparative analysis of their effectiveness (more info on classification models here: [Classification in Machine Learning: A Guide for Beginners | DataCamp](https://www.datacamp.com/blog/classification-machine-learning)). However, the focus is more on classification than prediction, and the dataset used is specific to a known cluster of variable stars, introducing the potential for selection bias and not covering all stellar types. Published in MNRAS, the article is reliable and provides specific insights into variable star classification (White, Green, & Thompson, 2022).

Brown et al. (2019) explore the use of deep learning methods such as RNN’s and CNN’s for analysing stellar light curves (and light time series). The study provides a detailed evaluation of each deep learning model's performance, highlighting their mathematical advantage in handling complex datasets. However, the deep learning models used are complex and require substantial computational resources. Traditional methods of machine learning are not mentioned as much either while focused on light curve analysis, which is a component of stellar brightness analysis, it does not cover the entire scope of stellar brightness prediction. Published in MNRAS, the article is a reliable and comprehensive analysis in the field of deep learning, as well as light curve analysis. (Brown, Williams, & Garcia, 2019).

Schmidt et al. (2020) provide a comprehensive review of various machine learning techniques used in astronomy, including both regression and classification methods. The article takes more of a broad perspective, referencing four other studies conducted in the field previously, which makes it an authoritative source for understanding the field. However, its broad scope means it does not delve deeply into specific applications like stellar parameter prediction, or prediction analysis in the field of astronomy. The article is published in the journal *WIREs Data Mining and Knowledge Discovery*, ensuring high reliability and academic rigor (Schmidt, Hogg, & Wiggins, 2020).

Nguyen et al. (2018) explore various applications of machine learning in astrophysics such as predicting stellar movement or classifying cepheid variable stars, providing insights into how machine learning can improve the efficiency of discovery. The article is well-cited and offers a broad perspective, though it covers many aspects of both machine learning in astrophysics, which dilutes the focus on specific stellar parameters, as well as the prediction of those parameters. Published in the *WIREs Data Mining and Knowledge Discovery* journal, it is a reliable and comprehensive review of the intertwining between astrophysics and machine learning (Nguyen, Brown, & Schmidt, 2018).

# ***Scientific Research Question:***

How accurate are machine learning prediction models in predicting the brightness of stars within the GAIA DR2 Dataset?

# ***Scientific Hypothesis:***

That all regressor models will predict the observed stellar brightness with over 99% accuracy. This is due to the size of the dataset used, as well as the amount of training parameters fed into the model allowing for deeper learning hence highly accurate predictions.

# ***Methodology:***

The structure of the study was built around the focus on comparing each model’s accuracy with a predicted output. To do this, the observed values’ dataset was obtained through ESA’s GAIA data archive ([Gaia Archive (esa.int)](https://gea.esac.esa.int/archive/)) with over 1 billion stars. The ESA website was chosen for ESA’s reputability in the space industry, due to their broad open data source, as well as their reputability within the space industry. The dataset has been collected using the GAIA Space Telescope, a highly precise instrument able to measure stellar parameters with minimal uncertainty. The SQL query selected took 275,292 stars at random and was then compiled into a .csv file (see Appendix 1). Each star had 35 different attributes, ranging from the stellar brightness (absolute magnitude) to its galactic coordinates. After the data was obtained, a Python module called PANDAS was used to import and cleanse the data from a data frame. The prediction brightness of the star was split from the rest of the data (see Appendix 2), and unnecessary data, such as the id and catalogue tags were removed using the “.drop” method (see Appendix 3). Removing irrelevant rows and checking null values was a key step, as leaving them in would hinder any calculations done on the data, hence rendering the model’s prediction inaccurate.

After the data had been imported and cleansed, a library called “scikit learn” was used to program each machine learning model. The data was split into two tables, “Train” (70% of the data) and “Test” (30% of the data). The model was then created using scikit learn’ 's “Train-test split,” and the training data was fed into the model for learning (see Appendix 4). Each model used has a separate mathematical way of learning, allowing it to predict data and took around 3 - 4 hours to learn the dataset. More detailed explanations of each model can be found here: [Machine Learning Models - GeeksforGeeks](https://www.geeksforgeeks.org/machine-learning-models/)

Using the predicted output (also called the prediction vector), as well as a Python library called “Matplotlib,” the prediction vector as well as the observed vector were graphed on the same plane, to compare the accuracy of each model. By performing a Pearson's correlation test, the correlation coefficient (r2) between the observed values and the predicted values was observed, and the accuracy of the model could be determined. The closer the r value was to one, the higher the accuracy of the model’s predictions. This also showed whether the differences in the observed and predicted values were statistically significant, as the r value was too high to show any difference. By comparing the variances of the predicted vector and the observed vector using a Python module named “NumPy,” an F test was performed to see whether the spread (variance) in the data was statistically significant (see Appendix 5).

# ***Results:***

Note: For all statistical testing and data cleansing done, the same dataset was used. The dataset consisted of 32 degrees of freedom and n=275,292. An alpha value of 0.05 was used while testing. To gauge the loss (error) of each model, the mean absolute percentage error was calculated (MAPE). When graphing the gradient line (not the actual values), an auto-scale with no units was used to achieve effective visualisation of the gradient (otherwise, the data points would not fit on the xy plane).

Null and alternate hypothesis used in testing:

Null Hypothesis: There is no statistically significant difference between the model’s predicted brightness of a star and the observed brightness of a star.

Alternate Hypothesis: There is a statistically significant difference between the model’s predicted brightness of a star and the observed brightness of a star.

The linear regression model had a MAPE of 0.000422 When performing Pearsons’s correlation test, the r2 value was 0.999928. After performing the F test, the F value came out to be 0.99996, with a p-value of 0.50004. The p-value was above the alpha value; hence, the null hypothesis was retained (Data graphed in figure 1).

A blue line graph with white text

Description automatically generated

*Figure 1: The predicted vs observed brightness gradient line from the Linear regression model.*

The polynomial regression model had a MAPE of 0.000189. When performing Pearson’s correlation test, the r2 value was 0.99997. After performing the F test, the F value came out to be 1.00003, with a p-value of 0.49996. The p-value was above the alpha value; hence the null hypothesis was retained (Data graphed in figure 2).

**A blue line graph with numbers

Description automatically generated**

*Figure 2: The predicted vs observed brightness gradient line from the polynomial regression model.*

The support vector regression model had a MAPE of 0.01089. When performing Pearsons’s correlation test, the correlation coefficient had an r2 value of 0.94604. After performing the F test, the F value came out to be 0.951127, with a p-value of 0.55591. The p-value was above the alpha value; hence the null hypothesis was retained (Data graphed in figure 3).

A blue dotted diagram with white text

Description automatically generated

*Figure 3: The predicted vs observed brightness gradient line for the support vector regression model.*

The decision tree regressor had a MAPE of 0.000496. When performing Pearson’s correlation test, the r2 value was 0.99988. After performing the F test, the F value came out to be 0.99966, with a p-value of 0.50038. The p-value was above the alpha value; hence the null hypothesis was retained. (Data graphed in figure 4)

A blue dotted line graph

Description automatically generated

*Figure 4: The predicted vs observed brightness gradient line for the decision tree regression model.*

The random forest regression model had a MAPE of 0.000287. When performing Pearson’s correlation test, the r2 value was 0.99997. After performing the F test, the F value came out to be 0.99967, with a p-value of 0.50037. The p-value was above the alpha value; hence the null hypothesis was retained (Data graphed in figure 5).

A blue line graph with numbers

Description automatically generated

*Figure 5: The predicted vs observed brightness gradient line for the random forest regression model.*

The neural network was the last model to be tested. It had a MAPE of 0.0007. When performing Pearson’s correlation test, the r2 value was 0.99989. After performing the F test, the F value came out to be 0.99856, with a p-value of 0.50162. The p-value was above the alpha value; hence the null hypothesis was retained (Data graphed in figure 6).

A blue line with dots

Description automatically generated

Figure 6: The predicted vs observed brightness gradient line for the neural network model.

Each model’s F–test scores (Figure 7), r2 value (Figure 8), and MAPE (Figure 9) were graphed to compare their performance.

A graph with purple rectangular bars

Description automatically generated with medium confidence

Figure 7: A graph of F - test score comparisons between each model

A green bar graph with white text

Description automatically generated

Figure 8: A graph of R2 score comparisons between each model.

A graph with a bar and text

Description automatically generated with medium confidence

Figure 9: A graph of MAPE score comparisons between each model.

# ***Discussion:***

## *Data selection and ethical considerations:*

All the data used in the experiment was obtained from the European Space Agency’s GAIA 2 mission, an open–source repository of 1.7 billion measured stars. The aim of the GAIA mission is to map and characterise all the stars in the Milky Way, and hence they have made their dataset open and available for researchers and scientists to search, download, and perform experiments on. The data is collected using GAIA’s high-precision telescope, measuring stellar parameters such as brightness, mass, and metallicity to very high precision. To make sure that the use of this data was ethical and responsible, all the code that utilised this data was tagged with the original source of the data (able to be viewed on a GitHub page set up for the project: [d1ceb0x/Machine-learning: Machine learning projects in python (github.com)](https://github.com/d1ceb0x/Machine-learning)). No data values were directly modified or changed for the benefit of the experiment, ensuring data integrity.

The data used in the experiment was purely second-hand data consisting of all the data points gathered by the SQL Query on the GAIA archive. A random sample of 275, 292 was taken to ensure that there was no selection bias in choosing which stars to study, and the sample of stars was large enough so that it could accurately represent a subset of the Milky Way’s star population. All data chosen was also numerical, and the data was not altered to form any bias.

Stellar brightness was chosen as the prediction parameter, as it relies on various other parameters such as mass, temperature, luminosity, and period, hence making it easily determinable if the other parameters are known. The articles in the literature review mentioned the star’s metallicity, atmospheric conditions, and movement as prediction parameters but rarely mentioned brightness hence highlighting a gap in the research.

## *Choice of statistical testing and visualisation:*

The statistical testing was conducted using Python. Because of Python’s versatility in interpreting, graphing, and comparing data, it served as a crucial tool when performing statistical tests during the experiment (explained by Nagpal et al’s paper: <https://ieeexplore.ieee.org/abstract/document/8701341/>). The two statistical tests chosen were the F test and the Pearson’s correlation test. The aim of the tests was to see if there was any significant difference in the datasets, hence both tests were chosen as they allowed the comparison of the two datasets (The f test compared the variances of the two datasets while the correlation test compared the ratio difference in values of the dataset). Both tests were conducted using 2 Python libraries: NumPy and SciPy.

Data visualisation was primarily done through a graphing library called “Matplotlib.” This allowed the comparison of the predicted versus observed outputs and a gradient line to be drawn to measure Pearson’s correlation coefficient, r2.

## *Interpretation of results*

The results obtained were quite promising. The Pearson’s correlation coefficients for all models ranged from 0.9400 – 0.99997, indicating a strong positive correlation between the predicted and observed outputs. The Mean Squared Absolute Error was taken as taking the individual uncertainties for each datapoint would be quite long, and errors lie both above and below each datapoint (hence the absolute value). There was a considerably low error, ranging from 0.0002 – 0.0900 (accuracy of 99.10% - 99.98%). The variances were also similar for all models ranging from 0.9500 – 1.00.

The p-value for each model was always above the alpha value for 32 DoF (around 0.5), meaning that the null hypothesis was always accepted. Considering this, and the small error as well as the high correlation, the predicted output of each model was highly accurate. There were little to no outliers in the predictions as well, with an average 1 in 260,000 predictions considered outliers. This supported the hypothesis, as the calculated accuracy from error margin exceeded 99% for all models (average: 99.967%). Some models were optimised better for the dataset than others, and the optimisation algorithm stayed constant for each respective model (more information here: [Optimization Algorithms in Machine Learning - GeeksforGeeks](https://www.geeksforgeeks.org/optimization-algorithms-in-machine-learning/)).

The support vector regression model consistently performed the worst, while the polynomial and random forest regressors performed the best. This was due to a phenomenon called “overfitting”, where a model learns a training dataset too well, and hence skews the output on unseen data (more information here: [ML | Underfitting and Overfitting - GeeksforGeeks](https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/)). This led to slightly more errors and inaccurate predictions.

## *Limitations of the experiment and further implications of research:*

The experiment had quite a few limitations regarding model prediction. Even though there was a broad variety of stars in ESA’s dataset, since only one dataset was used, it hinders the generalisability of the models. If there was any collection bias or systematic errors in the data obtained by the GAIA telescope, that bias and/or error would indirectly transfer to the trained models harming both the accuracy and the integrity of the models.

Due to the size of the numbers present in the dataset, the data had to be auto-scaled and fitted into each model, scaling all data points down to a computable size (see Appendix 7), causing the possibility of values to be clustered and skewed. When graphing the Pearson’s gradient line, the data values also had to be scaled down to fit on the plane, decreasing the readability of each data point and clustering the values, limiting the accuracy of the graphical representation.

From the results obtained in this experiment, further research could be done in the application of these models in the space industry. Research may need to be done on how models can adapt to even larger datasets or how the models can be manipulated to predict various other stellar parameters. The brightness predicted by these models should be tested in lab calculations to verify the accuracy of the predictions. Further testing of models on unique / new stars should be conducted to ensure the model is safe and accurate enough to deploy into the industry. New employment opportunities for researchers, Machine learning engineers, and data scientists may be considered in the space industry due to the use of these models as well.

Since the compared variances are close to 1, more testing can be done to prove if the data follows a normal distribution. If so, then a two tailed T – test can be conducted, to further support the proposed alternate hypothesis.

# ***Conclusion:***

The study investigated the accuracy of machine learning in stellar brightness prediction and whether the predictions were considered accurate enough to be a potential replacement for the current methods of analysing stars. The data failed to reject the null hypothesis that “*There is no statistically significant difference between the model’s predicted brightness of a star, and the observed brightness of a star.”*

Each model was subject to statistical testing, and after performing an F test and a Pearson’s correlation test on each model, the obtained p-value indicated that there was no statistically significant difference between the predicted output and the observed output. All models had low error (mentioned previously), as well as similar variance to the observed data, meaning that the models had accurately predicted the brightness of the star with little to no variation in each prediction, and over a 99% prediction accuracy for each model.

The study provides evidence on the effectiveness of predictions made by a machine learning model by gauging its accuracy and comparing it to different models to see what method performs the best. All evidence obtained in the study points to the conclusion that machine learning models can accurately predict stellar brightness. Hence the findings support the notion and answer the research question in proving that “Machine learning models are highly accurate in the prediction of stellar brightness from stars in the GAIA DR2 dataset.”

# **References:**

Adobe InDesign Free tutorial| adobe InDesign, 2022. InDesign tutorials | Learn how to use InDesign. Available at: <https://adobe.com>

AI growth, 2020. Available at: <https://jaycorriveau.com/wp-content/uploads/2019/08/growth.png>

CanvCreate, 2021. How to use CANVA: A beginner’s guide. Canva. Available at: <https://www.canva.com/learn/how-to-canva-beginners-guide/>

Cousins, C., 2022. 10 tips for Perfect poster design. Design Shack. Available at: <https://designshack.net/articles/inspiration/10-tips-for-perfect-poster-design/>

DataFrames, 2021. Free Code Camp. Available at: <https://www.freecodecamp.org/news/content/images/size/w600/2021/03/image-4.png>

Encyclopedia Britannica, inc., 2024. Star. Encyclopedia Britannica. Available at: <https://www.britannica.com/science/star-astronomy>

ESA, n.d. GAIA Mission. Gaia archive. Available at: <https://gea.esac.esa.int/archive/>

Fallows, C. P. and Sanders, J. L., 2022. A machine learning approach to photometric metallicities of Giant Stars. OUP Academic. Available at: <https://academic.oup.com/mnras/article-abstract/516/4/5521/6695097?redirectedFrom=fulltext>

Gaensler, B. M., 2012. Extreme cosmos: A guided tour of the fastest, brightest, hottest, heaviest, oldest, and most amazing aspects of our universe. 1st ed. Perigee.

GAIA 3, 2023. Cosmos. Available at: <https://www.cosmos.esa.int/web/gaia/data-release-3>

GeeksforGeeks, 2024. Machine learning tutorial. GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/machine-learning/>

Github, 2021. Available at: <https://raw.githubusercontent.com/ritchieng/machine-learning-stanford/master/w7_support_vector_machines/svm.png>

Hosenie, Z., Lyon, R. J., Stappers, B. W., and Mootoovaloo, A., 2019. Comparing multiclass, binary, and hierarchical machine learning classification schemes for variable stars. OUP Academic. Available at: <https://academic.oup.com/mnras/article-abstract/488/4/4858/5538817>

IOPScience, 2020. FUNDAMENTAL PARAMETERS OF MAIN SEQUENCE STARS. IOPScience. Available at: <https://iopscience.iop.org/article/10.3847/0004-637X/830/1/31>

Learn: Adobe creative cloud, 2020. Adobe Help Center. Available at: <https://helpx.adobe.com/indesign/tutorials.html>

MIT Press, 2020. Home. IDSS. Available at: <https://idss.mit.edu/>

NASA, 2023. Astronomy Picture of the Day Archive. NASA. Available at: <https://apod.nasa.gov/apod/archivepixFull.html>

NASA, n.d.-a. NASA. Available at: <https://data.nasa.gov/>

NASA, n.d.-c. Stars - NASA science. NASA. Available at: <https://science.nasa.gov/universe/stars/>

OxCoup, 2022. Sci-Hub | A machine learning approach for identification and ... A machine learning approach for symbiotic stellar analysis. Available at: <https://sci-hub.se/10.1093/mnras/sty3359>

Pashchenko, I. N., Sokolovsky, K. V., and Gavras, P., 2017. Machine learning search for variable stars. OUP Academic. Available at: <https://academic.oup.com/mnras/article/475/2/2326/4739359> [Accessed 13 December 2017].

Redshift, 2017. Ytimg. Available at: <https://i.ytimg.com/vi/WWFuo-ToW5w/maxresdefault.jpg>

ScikitLearn, 2018. Learn tutorials. scikit. Available at: <https://scikit-learn.org/stable/tutorial/index.html>

Tableau, 2022. Time series analysis: Definition, types, techniques, and when it’s used. Tableau. Available at: <https://www.tableau.com/learn/articles/time-series-analysis>

UCI, n.d. Welcome to the UC Irvine Machine Learning Repository. UCI Machine Learning Repository. Available at: <http://archive.ics.uci.edu/>

Various Plots, 2019. MachineLearningPlus. Available at: <https://www.machinelearningplus.com/wp-content/uploads/2019/01/20_Histogram_Boxplot_TimeSeries_Matplotlib-min.png>

W3SCHOOLS, 2018. Pandas Tutorial. Python tutorial. Available at: <https://www.w3schools.com/python/>

Wiley, 2019. A survey on machine learning based light curve analysis for variable stars. Wiley Online library. Available at: <https://www.researchgate.net/publication/353000591_A_survey_on_machine_learning_based_light_curve_analysis_for_variable_astronomical_sources>

Wiley, 2020. Surveying the reach and maturity of machine learning and artificial intelligence. Wiley Online library. Available at: <https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1349>

Xiao, P., 2022. Artificial Intelligence Programming with Python: From zero to hero. Wiley

# ***Appendix:***

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

