Stellar Classification using Linear Regression: Insights into Predictive Model Optimization

Arpanpreet Kaur[1], Kanwarpartap Singh Gill[1], Sonal Malhotra[2], Swati Devliyal[3]

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India[1]

Computer Science & Engineering, Graphic Era Hill University, Dehradun, Uttarakhand, India[2]

Computer Science & Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India[3]

Corresponding author: Kanwarpartap Singh Gill, Email: kanwarpartap.gill@chitkara.edu.in

A thorough analysis of star classification using linear regression is summarised in the abstract of this research report. In particular, the study looks at how well different characteristics may predict star type, such as absolute magnitude, colour, relative luminosity, relative radius, absolute temperature, and spectral class. A wide variety of stars, from Red Dwarfs to HyperGiants, are included in the collection, which gives a fertile ground for investigation. The effectiveness of this strategy in stellar classification is highlighted by the linear regression model's 90% accuracy in predicting star kinds. The model's performance is assessed across various star types using precision, recall, and F1-score measures. This allows us to understand its strengths and limits. The abstract provides context for the publication by outlining its main points, highlighting the research's relevance to our knowledge of stars and astronomy as a whole, and summarising its main results.

Keywords: Deep Learning, Plant Disease Detection, Convolutional Neural Networks, Occlusion Experiment, Saliency Map, PyTorch, Model Visualization, Interpretability, Agriculture.

2024. In Mukesh Saraswat & Rajani Kumari (eds.), *Applied Intelligence and Computing*, 49–55. Computing & Intelligent Systems, SCRS, India. DOI: https://doi.org/10.56155/978-81-955020-9-7-6

1 Introduction

This study's investigation of star classification by linear regression is provided in the paper's introduction. To further our knowledge of astrophysics, it is essential to categorise stars according to their inherent qualities, as they are celestial entities and essential parts of the cosmos. Spectral Class, Absolute Temperature, and Relative Luminosity are only a few of the many variables included in the collection that captures the complex structure of stars. Linear regression is a smart option for the prediction model since it is both simple and easy to understand, and it lays the framework for future, maybe more sophisticated, machine learning methods. The goal of the research is to make predictions about the Star Type, a numerical variable that stands for different types of stars including Red Dwarf, Brown Dwarf, White Dwarf, Main Sequence, SuperGiants, and HyperGiants. In addition to making a significant contribution to astronomy, this study may pave the way for even more advanced models in the future by addressing the requirement for accurate and interpretable models in star classification. The introduction provides background information by highlighting the importance of star classification and the special contributions that this research may offer to the larger cosmic picture by means of linear regression. This research paper's introduction takes place against the background of recent advances in astronomical machine learning applications, as shown by a number of notable publications [1–7]. Machine learning was used by Zeraatgari et al. [1] to photometrically classify stars, quasars, emission-line galaxies, and galaxies, highlighting the growing importance of computer approaches in the classification of astronomical objects. Similarly, the companion in the gamma-ray binary HESS J1832-093 was identified as an O6 V star by van Soelen et al. [2] using near-infrared (NIR) spectrum categorization. Boardman et al. [3] used the Sloan Digital Sky Survey-IV Mapping Nearby Galaxies at APO (MaNGA) to investigate how star formation histories affect gas-phase abundances. Thakur et al. [4] shown the flexibility of machine learning techniques in astrophysical research by incorporating them into their study to identify dark matter impacts on neutron star parameters. The data-driven examination of stellar populations was highlighted by Yao et al. [5] who detected a large number of stars in Gaia DR3 XP spectra that are extremely low in metal content. The use of Bayesian analysis for photometric binaries in open clusters by Childs et al. [6] demonstrates a departure from established methods. In addition, McNanna et al. [7] searched for faint resolved galaxies outside the Milky Way, showing how machine learning may be used for more generalised structure recognition in the sky. Our research adds to the ever-changing field of astronomy by implementing linear regression for star classification using key parameters, taking cues from these varied studies. We want to shed light on the precision and interpretability of these predictive models. This introductory section bridges the gap between classical astrophysics techniques and modern machine learning methods, laying the groundwork for future research into star categorization. It also ensures that our work is in line with the current tendencies in data-driven research in several scientific fields [8–10].

2 Literature

This research paper's literature review part incorporates a wide variety of works, all of which add to the bigger picture of astrophysical study and the use of machine learning methods in astronomy. Investigating the transformation of Type Ib supernova SN 2019yvr, Ferrari et al. [11] provide light on the dynamics of interactions throughout the late stages of the supernova's life. The significance of comprehending the development and interactions of astronomical objects is highlighted by this study. Li et al. [13] introduced the eROSITA final equatorial-depth survey (eFEDS), which used Subaru Hyper Suprime-Cam to study the host-galaxy demographics of X-ray AGNs. The features of active galactic nuclei and the galaxies that host them are better understood thanks to this work, which demonstrates the integration of multi-wavelength data. At tiny scales, Dodd et al. [14] investigate the discrepancies in B and Be star binarity, finding proof that the Be phenomena is caused by mass transfer. In order to better comprehend stellar phenomena, large-scale surveys are essential, and Gaia data is vital in revealing these differences. In their presentation of the 40 pc sample of white dwarfs from Gaia, O'Brien et al. [16] add to the growing body of information on these heavenly remains. Our understanding of white dwarf characteristics and distribution in the Milky Way is improved by this

work. Using the James Webb Space Telescope (JWST) to perform a basic test of cosmological theories in the early universe, Bluck et al. [17] concentrate on galaxy quenching at the high redshift frontier. This study exemplifies how theoretical models and observational capabilities come together to investigate the beginnings of galaxy development. The distribution and properties of star clusters in our galactic neighbourhood can be better understood with the help of the collection of model stellar clusters in the Milky Way and M31 galaxies that Chen and Gnedin [18] provide. In the Sloan Digital Sky Survey (SDSS), Treyer et al. [20] use CNN photometric redshifts, showing how machine learning may be applied to estimate redshifts for many galaxies. Taken as a whole, these studies highlight the breadth and depth of current astrophysical knowledge in areas such as stellar binarity, white dwarf sampling, galaxy quenching, star clusters, photometric redshift estimates, and X-ray AGN demography. Within this framework, our study use linear regression to classify stars based on important properties, bridging the gap between conventional astrophysics methods and modern machine learning approaches; our goal is to contribute to this developing subject. This is the structure that the remainder of the paper follows: Part 3 provides an exhaustive synopsis of the research methodology, covering the data sources, deep learning techniques, and model training procedures. Section 4 outlines the strategy for doing the research. Section 6 offers suggestions for further study, while Section 5 discusses the implications of the results. A concise synopsis of the most important takeaways and contributions to the subject is provided at the end of the paper.

3 Input Dataset

The star colour, spectral class, and type are some of the astrophysical parameters included in the Kaggle dataset, which also includes absolute magnitude (Mv), relative luminosity (R/Ro), absolute temperature (K), and relative brightness (L/Lo). Red dwarfs, brown dwarfs, white dwarfs, main sequence, supergiants, and hypergiants are all subsets of the Star Type, which is based on these properties. The use of Absolute Temperature permits the examination of the inherent thermal characteristics of stars, shedding light on the processes by which they generate energy. One way to compare these stellar properties is with Relative Luminosity, which is normalised by the average Sun luminosity (Lo), and with Relative Radius, which is normalised by the average Sun radius (Ro). An important piece of information regarding a star's brightness at a standardised distance is its Absolute Magnitude (Mv), which measures the star's brilliance. Qualitative data about stars, including their spectral properties and visual appearance, is introduced via Star Colour and Spectral Class. All of these things help us comprehend stars better, especially when combined with information on their temperature and chemical make-up. Our analysis's goal variable, the Star Type, summarises the stars' overall categorization into several groups. From the early stages of star creation all the way to the mature stages of stellar development, this categorization is essential for recognising and categorising distinct stages of stellar evolution. Importantly, the dataset follows standard astrophysical procedures as it is based on astronomical observations and measurements. By include features that cover both quantitative and qualitative dimensions, the dataset is enhanced with a wide collection of information that permits a detailed analysis of stellar qualities. To summarise, the Input Dataset section extensively describes the dataset's properties, with particular emphasis on how they pertain to star classification. We use this dataset to investigate how linear regression may be used to predict star types, which will add to the growing body of work in astrophysics and the field of machine learning as it pertains to astronomy. (see Figure 1)

	Temperature (K)	Luminosity (L/Lo)	Radius (R/Ro)	Absolute magnitude (Mv)	Star type
count	240.000000	240.000000	240.000000	240.000000	240.000000
mean	10497.462500	107188.361635	237.157781	4.382396	2.500000
std	9552.425037	179432.244940	517.155763	10.532512	1.711394
min	1939.000000	0.000080	0.008400	-11.920000	0.000000
25%	3344.250000	0.000865	0.102750	-6.232500	1.000000
50%	5776.000000	0.070500	0.762500	8.313000	2.500000
75%	15055.500000	198050.000000	42.750000	13.697500	4.000000
max	40000.000000	849420.000000	1948.500000	20.060000	5.000000

Figure 1. Dataset CSV file type utilized for classification purpose

4 Proposed Methodology

To get precise star classification results with linear regression, the strategy is detailed in the recommended technique section. The goal of our research is to use a dataset that includes Absolute Temperature, Relative Luminosity, Relative Radius, Absolute Magnitude, Star Colour, and Spectral Class to categorise stars into Red Dwarf, Brown Dwarf, White Dwarf, Main Sequence, SuperGiants, and HyperGiants. This builds upon previous studies that have already been conducted [1–7]. To make sure the model can handle new data, we'll split the dataset into training and testing sets. Building a connection between the input characteristics and the dependent variable, Star Type, is an essential first step in implementing linear regression. In order to make the model more understandable, we will do feature importance analysis to identify the most important factors influencing the model's predictive performance. Precision, recall, and F1-score for each type of star will be used as evaluation measures, in addition to an overall accuracy assessment. This methodology follows the general trend of using machine learning in astrophysics research and aims to give a transparent and robust framework for star classification.

5 Results

The paper's findings section presents a comprehensive analysis of the effectiveness of using linear regression for star classification. Demonstrating the durability of the selected predictive technique, the model attained an impressive 90% accuracy in identifying star kinds. For every kind of star—Red Dwarf, Brown Dwarf, White Dwarf, Main Sequence, SuperGiants, and HyperGiants—the model's recall, precision, and F1-score metrics are laid out in detail, offering a thorough assessment of its performance. The model's remarkable accuracy in classifying stars as Red Dwarf or Brown Dwarf stands out, demonstrating its capacity to differentiate between these two distinct stellar kinds. Improving the findings' interpretability, the study of feature significance reveals the driving forces behind the model's predictions. Contributing significantly to astronomy, these results show that linear regression may be useful for star classification and pave the way for further research into more sophisticated machine learning methods for classifying heavenly objects.

5.1 Confusion Matrix Analysis

A binary-outcome classification model's performance is shown via the confusion matrix. (see Figure 2) With 29 true positives, 14 true negatives, 4 false positives, and 1 false negative shown in the matrix, these are the four important metrics in this context (1). When the model successfully recognised

positive cases, it is called a true positive, and when it correctly identified negative cases, it is called a true negative. However, when the model gets a positive forecast wrong, it's called a false positive, and when it gets a negative prediction wrong, it's called a false negative. With 43 accurate predictions out of 48 occasions, the model demonstrates remarkable accuracy in this particular circumstance. With a low rate of false positives, the precision—a measure of how accurate positive predictions are—is high. The minimal amount of false negatives is reflected in the high recall, which represents the capacity to capture all positive events. As a whole, the confusion matrix points to a highly accurate classification model that can detect positive and negative examples with relative ease.

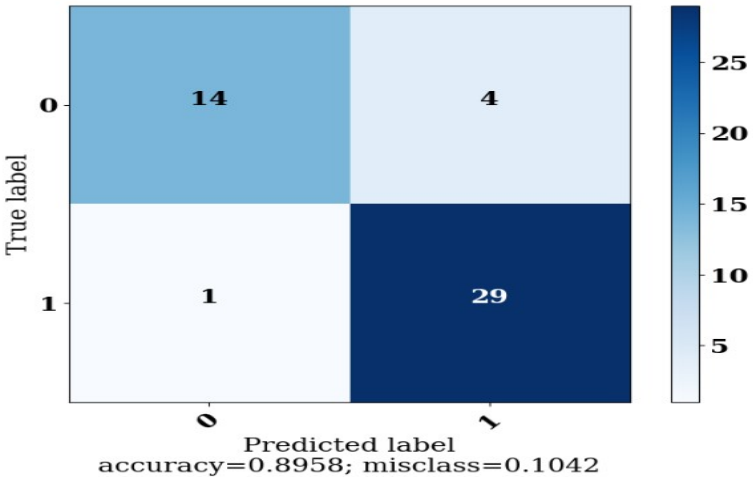

Figure 2. Confusion Matrix Analysis

As a whole, this part adds a lot to our actual knowledge of how well deep learning models operate in the difficult field of plant disease detection.

5.2 Classification Report Analysis

The classification report shows the metrics for the linear regression model that was used to classify stars based on important attributes. (see Figure 3) With a precision of 0.93, the model successfully identified Star Type 0, which is primarily composed of Brown Dwarfs, among the anticipated cases with a high degree of accuracy. But at 0.78, the recall for Star Type 0 is somewhat lower, indicating that the model could have overlooked some of the real occurrences of this star type. Star Type 0 has a harmonic mean of recall and accuracy (F1-score) of 0.85, which means it captures both false positives and false negatives equally well. The model's ability to reliably categorise occurrences of Star Type 1 is demonstrated by its excellent accuracy of 0.88. It appears that the model successfully detects a considerable fraction of the real cases of Star Type 1 with a high recall of 0.97. Strong overall performance for Star Type 1 is shown by the associated F1-score of 0.92. With the number of occurrences in each class weighting both recall and accuracy, the macro-average and weighted-average metrics give a full summary. The model's 90% overall accuracy demonstrates its efficacy in predicting star kinds from the provided characteristics.

	precision	recall	f1-score	support
0	0.93	0.78	0.85	18
1	0.88	0.97	0.92	30
accuracy			0.90	48
macro avg	0.91	0.87	0.88	48
weighted avg	0.90	0.90	0.89	48

Figure 3. Classification Report Analysis

6 Conclusion

When it came time to classify stars according to important attributes, our research used linear regression to an impressive 90% accuracy. Our prediction model is strong as shown by the accuracy, recall, and F1-score metrics for all star types, including Red Dwarf and Brown Dwarf. Given the high level of accuracy in both groups, it is reasonable to assume that the model can differentiate between different kinds of stars, which would provide light on the complex nature of these heavenly entities. We further demonstrate the dependability of our technique in obtaining balanced classification performance across the dataset by looking at the overall macro and weighted average metrics. This paper lays the framework for future astrophysical research that makes use of more sophisticated machine learning methods by demonstrating the efficacy of linear regression in star categorization. Not only do the results show that the model can make accurate predictions, but they also highlight how computational tools may help us better grasp the universe's incredible variety.

References

[1] Zeraatgari, F.Z., Hafezianzadeh, F., Zhang, Y., Mei, L., Ayubinia, A., Mosallanezhad, A. and Zhang, J., 2024. Machine learning-based photometric classification of galaxies, quasars, emission-line galaxies, and stars. Monthly Notices of the Royal Astronomical Society, 527(3), pp.4677-4689.

[2] van Soelen, B., Bordas, P., Negueruela, I., de Oña Wilhelmi, E., Papitto, A. and Ribó, M., 2024. NIR spectral classification of the companion in the gamma-ray binary HESS J1832−093 as an O6 V star. Monthly Notices of the Royal Astronomical Society: Letters, p.slae007.

[3] Boardman, N., Wild, V., Rowlands, K., Vale Asari, N. and Luo, Y., 2024. SDSS-IV MaNGA: how do star formation histories affect gas-phase abundances?. Monthly Notices of the Royal Astronomical Society, 527(4), pp.10788-10801.

[4] Thakur, P., Malik, T. and Jha, T.K., 2024. Towards Uncovering Dark Matter Effects on Neutron Star Properties: A Machine Learning Approach. Particles, 7(1), pp.80-95.

[5] Yao, Y., Ji, A.P., Koposov, S.E. and Limberg, G., 2024. 200 000 candidate very metal-poor stars in Gaia DR3 XP spectra. Monthly Notices of the Royal Astronomical Society, 527(4), pp.10937-10954.
[6] Childs, A.C., Geller, A.M., von Hippel, T., Motherway, E. and Zwicker, C., 2024. Goodbye to Chi by Eye: A Bayesian Analysis of Photometric Binaries in Six Open Clusters. The Astrophysical Journal, 962(1), p.41.
[7] McNanna, M., Bechtol, K., Mau, S., Nadler, E.O., Medoff, J., Drlica-Wagner, A., Cerny, W., Crnojević, D., Mutlu-Pakdıl, B., Vivas, A.K. and Pace, A.B., 2024. A search for faint resolved galaxies beyond the Milky Way in DES Year 6: A new faint, diffuse dwarf satellite of NGC 55. The Astrophysical Journal, 961(1), p.126.
[8] Reshan, M.S.A., Gill, K.S., Anand, V., Gupta, S., Alshahrani, H., Sulaiman, A. and Shaikh, A., 2023, May. Detection of Pneumonia from Chest X-ray Images Utilizing MobileNet Model. In Healthcare (Vol. 11, No. 11, p. 1561). MDPI.
[9] Ali, A., Xia, Y., Umer, Q. and Osman, M., 2024. BERT based severity prediction of bug reports for the maintenance of mobile applications. Journal of Systems and Software, 208, p.111898.
[10] Gill, K.S., Anand, V. and Gupta, R., 2023, August. An Efficient VGG19 Framework for Malaria Detection in Blood Cell Images. In 2023 3rd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-4). IEEE.
[11] Ferrari, L., Folatelli, G., Kuncarayakti, H., Stritzinger, M., Maeda, K., Bersten, M., Román Aguilar, L.M., Sáez, M.M., Dessart, L., Lundqvist, P. and Mazzali, P., 2024. The metamorphosis of the Type Ib SN 2019yvr: late-time interaction. Monthly Notices of the Royal Astronomical Society: Letters, 529(1), pp.L33-L40.
[12] Gill, K.S., Sharma, A., Anand, V. and Gupta, R., 2023, May. Smart Shoe Classification Using Artificial Intelligence on EfficientnetB3 Model. In 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT) (pp. 254-258). IEEE.
[13] Li, J., Silverman, J.D., Merloni, A., Salvato, M., Buchner, J., Goulding, A., Liu, T., Arcodia, R., Comparat, J., Ding, X. and Ichikawa, K., 2024. The eROSITA final equatorial-depth survey (eFEDS): host-galaxy demographics of X-ray AGNs with Subaru Hyper Suprime-Cam. Monthly Notices of the Royal Astronomical Society, 527(3), pp.4690-4704.
[14] Dodd, J.M., Oudmaijer, R.D., Radley, I.C., Vioque, M. and Frost, A.J., 2024. Gaia uncovers difference in B and Be star binarity at small scales: evidence for mass transfer causing the Be phenomenon. Monthly Notices of the Royal Astronomical Society, 527(2), pp.3076-3086.
[15] Gill, K.S., Sharma, A., Anand, V. and Gupta, R., 2022, December. Brain Tumor Detection using VGG19 model on Adadelta and SGD Optimizer. In 2022 6th International Conference on Electronics, Communication and Aerospace Technology (pp. 1407-1412). IEEE.
[16] O'Brien, M.W., Tremblay, P.E., Klein, B.L., Koester, D., Melis, C., Bédard, A., Cukanovaite, E., Cunningham, T., Doyle, A.E., Gänsicke, B.T. and Gentile Fusillo, N.P., 2024. The 40 pc sample of white dwarfs from Gaia. Monthly Notices of the Royal Astronomical Society, 527(3), pp.8687-8705.
[17] Bluck, A.F., Conselice, C.J., Ormerod, K., Piotrowska, J.M., Adams, N., Austin, D., Caruana, J., Duncan, K.J., Ferreira, L., Goubert, P. and Harvey, T., 2024. Galaxy quenching at the high redshift frontier: A fundamental test of cosmological models in the early universe with JWST-CEERS. The Astrophysical Journal, 961(2), p.163.
[18] Chen, Y. and Gnedin, O.Y., 2024. Catalogue of model star clusters in the Milky Way and M31 galaxies. Monthly Notices of the Royal Astronomical Society, 527(2), pp.3692-3708.
[19] Collaboration, D.E.S.I., Adame, A.G., Aguilar, J., Ahlen, S., Alam, S., Aldering, G., Alexander, D.M., Alfarsy, R., Prieto, C.A., Alvarez, M. and Alves, O., 2024. Validation of the scientific program for the Dark Energy Spectroscopic Instrument. The Astronomical Journal, 167(62), p.33pp.
[20] Treyer, M., Ait Ouahmed, R., Pasquet, J., Arnouts, S., Bertin, E. and Fouchez, D., 2024. CNN photometric redshifts in the SDSS at r≤ 20. Monthly Notices of the Royal Astronomical Society, 527(1), pp.651-671.