DRAFT VERSION MAY 9, 2022
Typeset using LaTeX **preprint** style in AASTeX63

Predicting Effective Temperatures of O and B Stars using Artificial Neural Networks

NIKKO J. CLERI[1,2]

[1]*Department of Physics and Astronomy, Texas A&M University, College Station, TX, 77843-4242 USA*
[2]*George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, Texas A&M University, College Station, TX, 77843-4242 USA*

ABSTRACT

In this work, we use artificial neural network regression to determine effective temperatures for 959 O and B type stars in the Gaia Early Data Release 3. We train our neural network on a sample of 1050 O and B stars from Gaia EDR3 with known effective temperatures from the Transiting Exoplanet Survey Satellite using 20 numerical predictors. Our validation shows that the neural network model gives reasonable population statistics for effective temperature, but the given data and sparsity of reliable predictors does not allow for effective prediction of temperatures for individual stars.

1. INTRODUCTION

In astrophysics, data are being produced by current generation observatories at a rate greater than the growth of the storage supply. In the soon-to-be generation of telescopes including the *Vera C. Rubin Observatory* (*VRO*), *Nancy Grace Roman Space Telescope* (*NGRST*), and the recently-launched *James Webb Space Telescope* (*JWST*), the production of data will greatly exceed the on-board storage capacity. These observatories will require some means of computationally efficient data preprocessing to avoid losing valuable information.

Machine learning methods for data processing are rising in popularity in astronomy research. One of the most common widely studied problems in astrophysics that has benefited greatly from the advent of machine learning is the classification of galaxy morphologies, studied using several different machine learning algorithms: feedforward neural networks, two-hidden-layer ANNs, random forests and decision trees, along with SVM, and K-means clustering and agglomerative hierarchical clustering (Storrie-Lombardi et al. 1992; Banerji et al. 2010; Gauci et al. 2010; Sreejith et al. 2018; Reza 2021).

Artificial neural networks (ANNs) have been used to classify stellar spectra as far back as 1998 (Singh et al. 1998), determining photometric redshifts (Firth et al. 2003; Way & Klose 2012), and other applications. Extremely randomized trees (ETs) have been used to estimates physical parameters of stellar atmospheres (Zhang et al. 2019) and in redshift estimation of nearby galaxies and distant quasars (Reza & Haque 2020). Support-vector machines (SVMs) have been used to classify spectral subclasses (Bu et al. 2014).

The remainder of this work is as follows. Section 2 discusses the data and sample selection for this work. Section 3 describes the model used for our neural network. Section 4 shows the results of our analysis. Section 5 discusses the implications of TensorFlow in the evolving picture of data-driven sciences.

2. DATA

Our data come from the Gaia Early Data Release 3 which gives stellar parameters for 2009 O and B type stars used in this work (Gaia Collaboration et al. 2016, 2021). We divide this sample into two subsamples: objects with effective temperatures from the Transiting Exoplanet Survey Satellite (TESS, Ricker et al. (2015)) catalogs (1050 objects), and those with no measured effective temperatures (959 objects). These data are heavily lopsided toward B stars, with O stars only making up $\approx 2.5\%$ of the sample.

Several of the parameters given in the Gaia EDR3 catalog are not suitable for the neural network analysis. Several of these, including mass and radius, are too sparsely supplied for our model to use them with any level of efficacy. We reduce our sample to only objects which have all of the 17 predictors which are shown in Table 1.

3. DESCRIPTION OF NEURAL NETWORK MODEL

Here we describe the neural network model used in this analysis. In general, we consider the inner workings of an artificial neural network (ANN) regression to be a black box. The input data is normalized and fed through the hidden layers of the neural network to achieve an output prediction.

We use neural network models from TensorFlow and the Keras algorithm library (Abadi et al. 2015; Chollet et al. 2015). We normalize our data to be usable by the neural networks using min-max scaling from `scikit-learn` preprocessing, which shifts and scales inputs into a distribution centered about 0 with a standard deviation of 1 (Pedregosa et al. 2011). The normalization improves the accuracy and computational efficiency of the neural network training (Sola & Sevilla 1997).

Table 1. Descriptions of Predictors from Gaia EDR3 Catalog

Predictor	Description
ra_edr3	Gaia EDR3 right ascension
dec_edr3	Gaia EDR3 declination
parallax_edr3	Gaia EDR3 parallax
pmra	RA proper motion
pmdec	Dec proper motion
phot_g_mean_mag	G band mean magnitude
Separation	Distance between matched objects along a great circle
Vmag	V band magnitude
e_Vmag	V band magnitude uncertainty
Gmag	G band magnitude
e_Gmag	V band magnitude uncertainty
Tmag	TESS I_c band magnitude
e_Tmag	TESS I_c band magnitude uncertainty
Dist	Distance from observer
s_Dist	Statistical uncertainty in distance
E(B-V)	Reddening
s_E(B-V)	Statistical uncertainty in $E(B-V)$
APP_MAG	Apparent magnitude
BV_COLOR_1	$B-V$ color
UB_COLOR_1	$U-B$ color

The important tunable hyperparameters of the neural network models include the activation function, the depth (number of layers), optimization function, loss function, and dimensionality of output layer, among others. This suite of tunable parameters allows for great amounts of customization and specialization for a wide variety of problems. Table 1 shows the hyperparameters used for our neural network model.

4. RESULTS

In this section we show the results of our neural network predictions. We divide our sample of 1050 stars with known effective temperature into training (85%) and validation (15%) subsamples. We train for 200 epochs to minimize the mean squared error. We show the loss function in Figure 2. Our model converges to minimize the loss function at approximately 100 epochs.

Figure 3 shows the distribution of the known effective temperatures for 1050 objects and the predictions for the 959 objects which do not have known effective temperatures. We note that the population statistics are reasonable compared to the population of known effective temperatures.

Figure 4 shows the relationship of known effective temperatures to predicted effective temperatures for the objects in our validation set. Although Figure 3 implies that the population statistics for the neural network predictions are reasonable, Figure 4 shows no clear trend with the one-to-one relationship. This implies that the prediction of the effective temperature for an individual star is not reliable in this analysis.

We show this further in Figure 5, which shows the distribution of the absolute deviation of effective temperatures between known and predicted in the testing sample. We find a median absolute deviation of 1566 K, which corroborates the failure of our model to accurately predict individual temperatures.

As described earlier, the neural network behaves effectively as a black box. We cannot use the results of the neural network in Figure 1 to draw any conclusions other than the predictive efficacy of the model. We conclude that this neural network is a fast and efficient means of prediction for this problem, but offers nothing in terms of interpretation.

5. CONCLUSIONS

In this work, we have used a sample of 1050 O and B stars from the Gaia EDR3 catalog with known effective temperatures to predict the effective temperatures of 959 other stars. To do this, we used an artificial neural network with 20 predictors, trained on 85% of the 1050 stars in the known population, validated on the other 15% of the known population, and tested on the 959 stars without known effective temperatures. We summarize our results as follows:

Table 1. Hyperparameters used in Neural Network Analysis

Parameter Description	Used Hyperparameter
Optimization Function	Adam Algorithm (Kingma & Ba 2014)
Loss Function	Mean Squared Error
Output Dimensionality	100
Number of Epochs	200

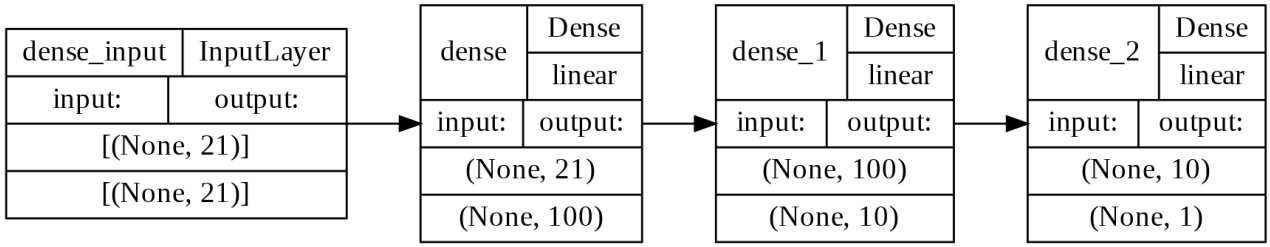

Figure 1. TensorFlow schematic of the neural network used for our AGN classification problem. The 17 input predictors are normalized in the process described in 3. The normalized inputs are then passed through the hidden layers of the neural network which result in the output layer.

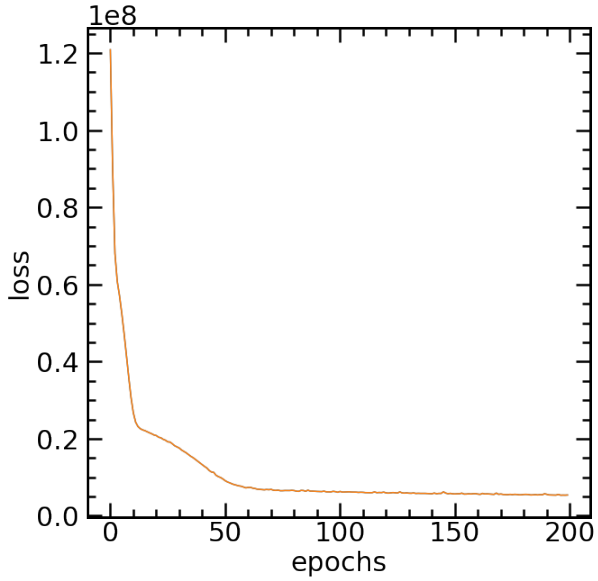

Figure 2. Loss as a function of epochs for our neural network model. Our loss is modeled by the mean squared error. Our model converges to minimize the loss at approximately 100 epochs.

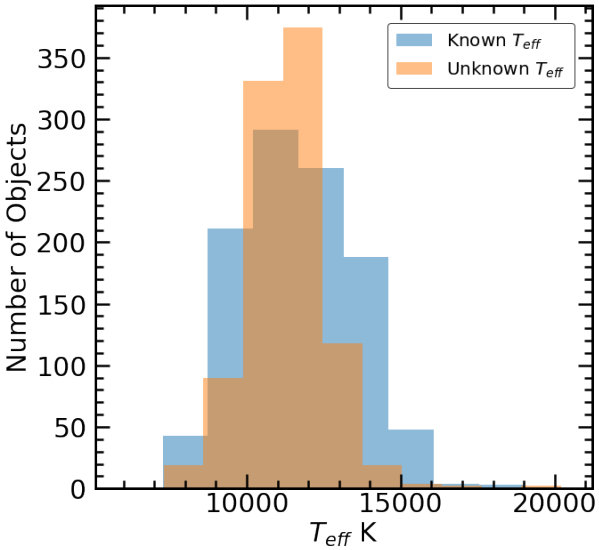

Figure 3. Histogram of the distribution of effective temperatures for the known sample (blue) and unknown sample (orange). The unknown sample gives a reasonable population distribution of temperatures compared to the known distribution. However, we argue that the prediction for an individual star is not necessarily reliable (see Figure 4)

- We find that the distribution of the population of effective temperatures for the testing sample is reasonably consistent with the effective temperatures of the known sample (See Figure 3).
- We find that our model fails to accurately predict the effective temperature of individual stars, in spite of the previous point. We attribute this to several of the very relevant predictors (mass, radius, etc.) being very sparsely available in the testing data, and as such were excluded from this analysis. The median absolute deviation of known versus predicted effective temperatures for our test sample is 1566 K.

In spite of the failure of our neural network model to accurately predict the effective temperatures of individual stars, we argue that this is likely only due to incomplete data and relatively small training sets. In future surveys with large-

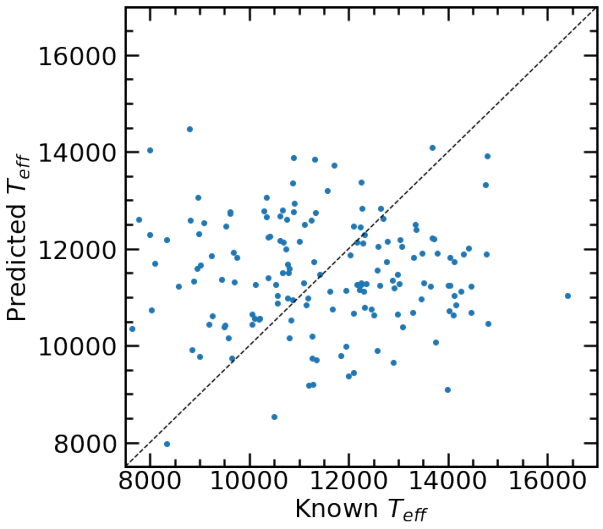

Figure 4. Predicted effective temperature versus known effective temperature for 158 objects in our validation sample. We show there is a clear scatter and no visible trend about the one-to-one line (black dashed), which indicates that the prediction for an individual star's temperature from this neural network model is not reliable.

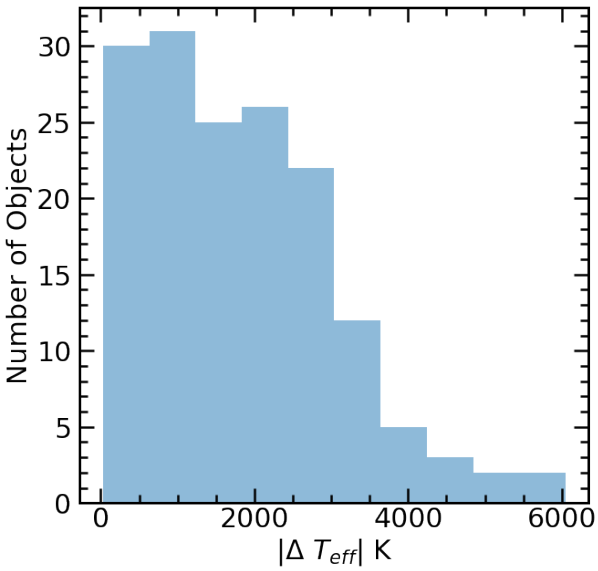

Figure 5. Histogram of the absolute deviations of the known and predicted effective temperatures for each star in our testing set. We have a median absolute deviation of 1566 K, further implying that this model with these data does not reliably predict individual temperatures.

number statistics, machine learning analyses will become increasingly necessary to avoid loss of information to slow and computationally inefficient preprocessing.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, , , software available from tensorflow.org. https://www.tensorflow.org/

Banerji, M., Lahav, O., Lintott, C. J., et al. 2010, MNRAS, 406, 342

Bu, Y., Chen, F., & Pan, J. 2014, NewA, 28, 35

Chollet, F., et al. 2015, Keras, https://keras.io, ,

Firth, A. E., Lahav, O., & Somerville, R. S. 2003, Monthly Notices of the Royal Astronomical Society, 339, 1195. https://doi.org/10.1046%2Fj.1365-8711.2003.06271.x

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, A&A, 649, A1

Gauci, A., Zarb Adami, K., & Abela, J. 2010, arXiv e-prints, arXiv:1005.0390

Kingma, D. P., & Ba, J. 2014, Adam: A Method for Stochastic Optimization, arXiv, doi:10.48550/ARXIV.1412.6980. https://arxiv.org/abs/1412.6980

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Reza, M. 2021, Astronomy and Computing, 37, 100492

Reza, M., & Haque, M. A. 2020, Ap&SS, 365, 50

Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003

Singh, H. P., Gulati, R. K., & Gupta, R. 1998, MNRAS, 295, 312

Sola, J., & Sevilla, J. 1997, IEEE Transactions on Nuclear Science, 44, 1464

Sreejith, S., Pereverzyev, Sergiy, J., Kelvin, L. S., et al. 2018, MNRAS, 474, 5232

Storrie-Lombardi, M. C., Lahav, O., Sodre, L., J., & Storrie-Lombardi, L. J. 1992, MNRAS, 259, 8P

Way, M. J., & Klose, C. D. 2012, PASP, 124, 274

Zhang, Y., Tu, Y., Zhao, Y., & Tian, H. 2019, in Astronomical Society of the Pacific Conference Series, Vol. 521, Astronomical Data Analysis Software and Systems XXVI, ed. M. Molinaro, K. Shortridge, & F. Pasian, 417