DEEP LEARNING FOR HARVARD SPECTRAL STAR CLASSIFICATION

Sonnet Salice
sonnet.salice@mail.utoronto.ca

Daniel Gonzalez
daniel.gonzalez@mail.utoronto.ca

Sophie Zheng
sophie.zheng@mail.utoronto.ca

ABSTRACT

Stellar classification is a cornerstone of modern astronomy, providing critical insights into stellar evolution, composition, and lifecycle. Manual classification, however, is time-intensive and impractical for analyzing massive datasets like the Gaia Data Release 3, which includes over 600,000 stars with diverse photometric features. This project employs Artificial Neural Networks (ANNs) to automate spectral classification into seven classes (O, B, A, F, G, K, M) based on four key features: Gmag, BPmag, RPmag, and Lum-Flame. The ANN achieved an accuracy of 84% on unseen test data, surpassing the baseline SVM model by 14%. Preprocessing techniques such as SMOTE for handling class imbalance and MinMax scaling were crucial for performance improvements. The model's effectiveness is further validated through a Hertzsprung-Russell (HR) diagram generated from predictions, which closely resembles real stellar distributions. While the model performs well overall, challenges persist in distinguishing overlapping classes, highlighting areas for future improvement. This project demonstrates the potential of deep learning to revolutionize stellar classification and support large-scale astronomical research.

1 INTRODUCTION

This project aims to classify stars based on their spectral characteristics, specifically photometric data (such as G, BP, and RP magnitudes) from the Gaia dataset, into spectral classes (O, B, A, F, G, K, and M types).

Accurately classifying stars provides insight into their age, composition, and temperature, contributing to our understanding of stellar evolution and the distribution of star types across the galaxy making it a key process in humanity's progression in astronomy and astrophysics. However, traditional methods for star classification rely on manual analysis or complex algorithms, which are often time intensive. A machine learning-based solution can speed up the classification process, making it feasible to analyze massive datasets from modern sky surveys like Gaia. Deep learning is particularly useful for this task due to its ability to learn complex, non-linear relationships within data.

The Gaia dataset provides numerous photometric features (e.g., magnitudes, colors), and neural networks can effectively process these continuous variables. By feeding in 4 input values (Gmag, BPmag, RPmag, Lum-Flame), the network can be trained to identify intricate patterns associated with each spectral class.

2 BACKGROUND AND RELATED WORK

Star classification has been a central focus of astronomy, with contributions ranging from manual classification systems to advanced computational approaches. Here we discuss five key works that provide the foundation and context for our project, highlighting their methodologies and relevance:

• *Morgan et al. (1978) (1)* - The MK (Morgan-Keenan) system refined the earlier Harvard spectral classification by incorporating luminosity classes, standardizing how stars are categorized. This framework remains foundational, providing the basis for modern classification systems that combine photometric and spectroscopic data.

• Gray and Corbally (2009) (2) - Building on the MK system, this study introduced luminosity subclasses, enhancing precision in categorizing stars with subtle spectral differences. This demonstrates the value of refining classification granularity—a principle we adopt in distinguishing overlapping spectral types (e.g., A and F stars) in our machine learning approach.

• Bailer-Jones et al. (2013) (3) - This study used machine learning techniques to classify stars using Gaia data, showing the feasibility of automated classification for large datasets. It highlights the potential of leveraging features like photometric magnitudes for accurate predictions, which aligns with our project's use of Gaia-derived data (Gmag, BPmag, RPmag, Lum-Flame).

• Fabbro et al. (2018) (4) - Demonstrating the applicability of Convolutional Neural Networks (CNNs) to astronomical spectra, this study revealed that deep learning models can effectively capture complex spectral patterns. While our work focuses on simpler features (photometric magnitudes and luminosity), this shows how neural networks can outperform traditional methods for high-dimensional data.

• *Mowlavi et al. (2021) (5)* - Using Gaia DR2 data, this study combined Support Vector Machines (SVM) and deep learning models for stellar classification. It emphasizes the importance of preprocessing (e.g., handling class imbalances and scaling features), which directly informs our use of SMOTE and MinMax scaling for data preparation.

These works illustrate the evolution of star classification, from manual to computational techniques. By building on these foundations, our project aims to extend the use of neural networks for automated stellar classification, focusing on accessible photometric features rather than full spectra. Unlike prior works that rely heavily on spectroscopic data, our approach offers a practical alternative by using easily obtainable features, potentially enabling faster and more scalable classification for large datasets.

3 The Model/Figure

Below is a basic idea of our classification model. Further details on the architecture of the ANN is given in the Architecture section of the report. We chose an ANN for this as the input features we used were all numerical and an ANN was best to find patterns in such data.

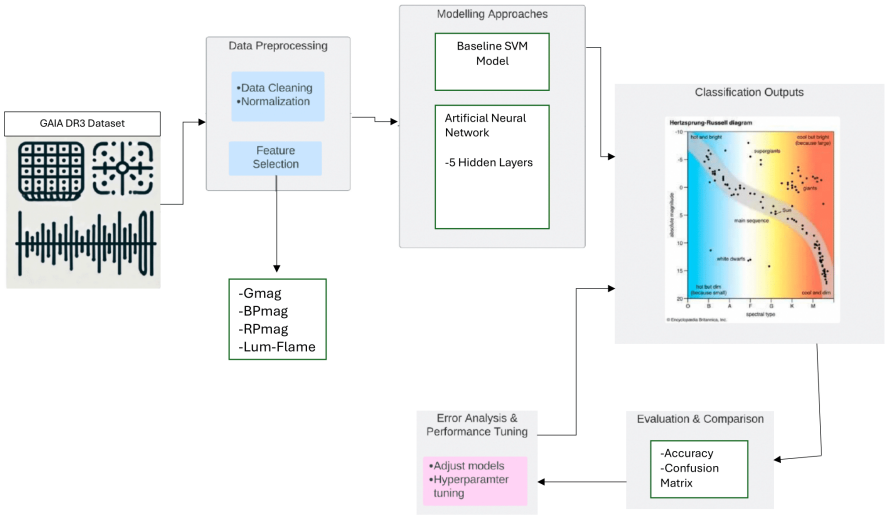

Figure 1: Deep Learning for Harvard Spectral Classification Model Illustration

4 DATA PROCESSING

Below we outline the Dataset selected and Data processing of our model.

4.1 DATASET SELECTION

The selected dataset is the Gaia Stars Dataset from DR3, accessible at `https://www.kaggle.com/datasets/realkiller69/gaia-stars-dataset-from-dr3-data-release-3`. This dataset is beneficial for our project for several reasons:

- **Large Dataset Size**: Contains 626,016 instances, providing a robust foundation for training deep learning models that require extensive data.
- **Diverse Spectral Classes**: Features around 100,000 instances for each spectral class (B, A, F, G, K, M) and 26,016 for O-type stars, ensuring balanced representation.
- **Rich Feature Set**: Includes up to 50 columns with valuable astrophysical parameters, essential for accurately classifying stars. Out of which the inputs we need are easy to use and there arent many rows with missing data for those inputs.
- **Directly Relevant Classification**: Contains the spectral class label (SpType-ELS) needed for the primary objective of automating stellar classification.
- **Strong Data Sources**: Sourced from reputable astronomical databases, ensuring high-quality and reliable data.
- **Potential for Insights**: Enables visualization of classifications through Hertzsprung-Russell diagrams.

4.2 DATA PROCESSING STEPS

1. **Loading the Dataset**: The Gaia Stars Dataset was loaded from Google Drive, and the first few rows were displayed to verify correctness.
2. **Dataset Overview**:
   - Used `df.info()` to get a summary of the dataset structure, revealing 626,016 instances with 50 columns.
   - Used `df.describe()` for statistical details of numerical columns.
   - Checked for missing values with `df.isnull().sum()` and found that several columns had missing data, particularly GRVSmag and RV.
3. **Data Cleaning**:
   - Rows with missing critical data (SpType-ELS) were dropped, leading to the removal of a total of 14,723 rows.
   - Entire columns with too many missing values (e.g., GRVSmag and RV) were removed.
   - The cleaned dataset now consists of 611,293 instances.
   - Example of a cleaned training sample:

```
RA_ICRS: 123.456
DE_ICRS: -45.678
Gmag: 13.9
BPmag: 14.6667
RPmag: 18.7
SpType-ELS: G
```

4. **Feature Selection**:
   - Key features for classification were selected: Gmag, BPmag, RPmag, Lum-Flame
   - The target variable was defined as SpType-ELS.
   - Class distribution statistics after cleaning:

```
Class Distribution:
O: 26,016
B: 100,000
A: 100,000
F: 100,000
G: 100,000
K: 100,000
M: 100,000
```

5. **Data Scaling**: Features were scaled using MinMaxScaler to normalize the data, which is crucial for many machine learning algorithms.
6. **Encoding Target Variable**: The target variable was encoded using LabelEncoder to convert string labels into numerical format.
7. **Train-Validation-Test Split**: The dataset was split into training (70%), validation (15%), and testing (15%) sets using `train_test_split`. This distribution allows for robust training while ensuring proper validation and testing.
8. **Handling Missing Values**: Checked for any remaining NaN values in the training set, and missing values in `X_train` were imputed using the mean strategy with SimpleImputer.
9. **Balancing the Dataset**: SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the training data, addressing potential class imbalance issues.
10. **Class Distribution**: The distribution of classes in the balanced training set was printed for verification, confirming that the application of SMOTE successfully equalized the instances across classes.

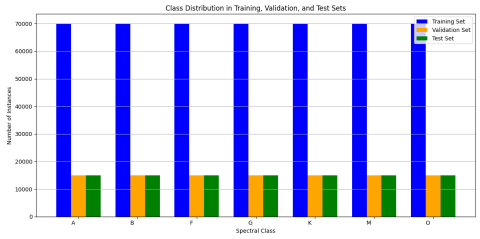

Figure 2: Class Distribution in Training, Validation, and Test Sets

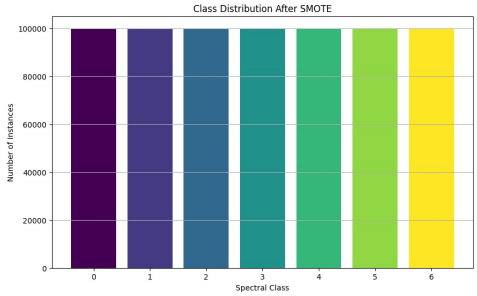

Figure 3: Class Distribution After SMOTE.

5 ARCHITECTURE

Model Description: We used an ANN for classifying the stars. The architecture of our ANN includes 4 input nodes, 5 hidden layers with up to 256 neurons, and an output layer with 7 nodes—each representing a spectral type. Our training parameters included an initial learning rate of 0.01, a batch size of 1,000, and 200 epochs. There is a Leaky ReLU activation between each layer, Cross Entropy Loss function, an Adam optimizer, a step-based learning rate scheduler and a weighted loss function.

Input: 4 spectral features from Gaia (Gmag, BPmag, RPmag, Lum-Flame).

- **Gmag**: Average brightness in the green band (mid-range wavelengths).

- **BPmag**: Brightness in the blue band (shorter wavelengths).

- **RPmag**: Brightness in the red band (longer wavelengths).

- **Lum-Flame**: Luminosity of star relative to the Sun.

Output: Automated classifications of stars into spectral types (O, B, A, F, G, K, M).

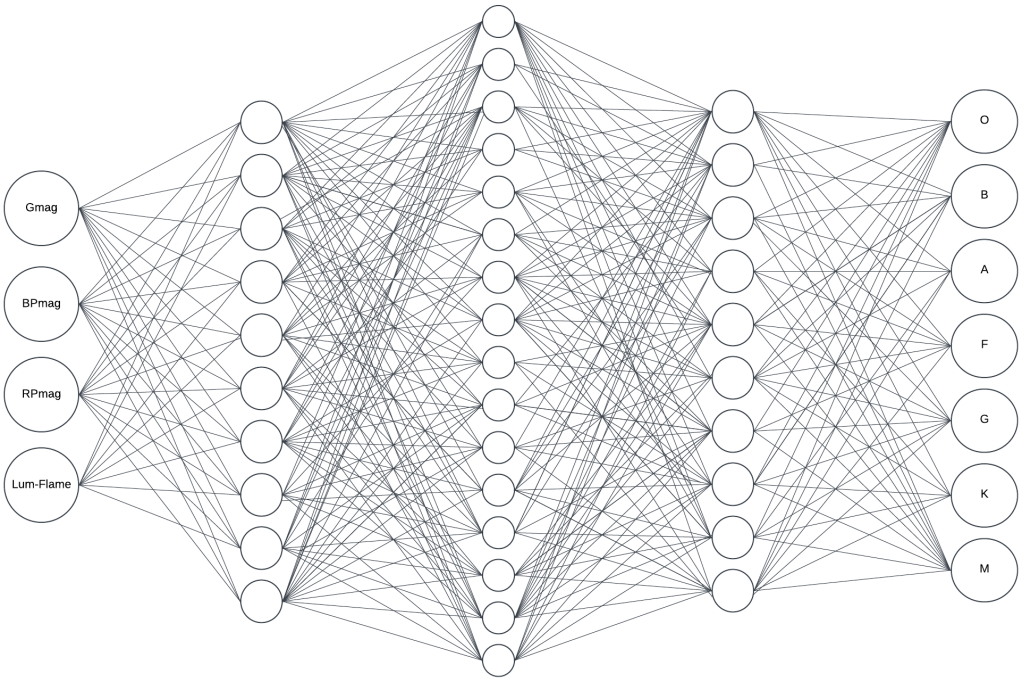

Figure 4: ANN Model Illustration (actual model has 5 hidden layers with 64, 128, 256, 128, 64 neurons each)

6 Baseline Model

We used a Support Vector Machine (SVM) for the baseline model. SVMs are commonly used for classification tasks, and they work well with smaller datasets or a subset of a larger dataset. This choice was made because SVMs are relatively simple to implement and interpret while providing reliable results without extensive tuning, making them suitable for assessing the feasibility of the primary model.

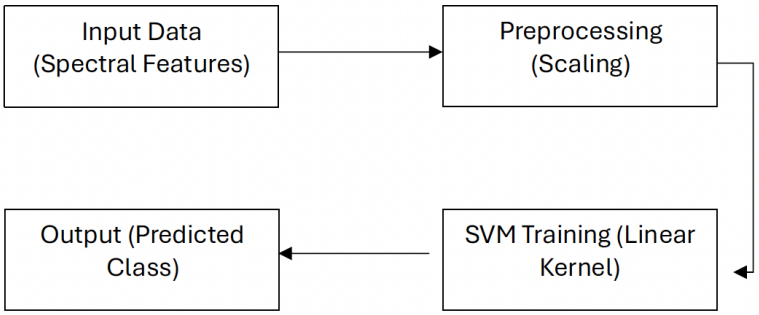

Figure 5: Workflow of Baseline SVM Model

7 Baseline Model Results

The confusion matrix (Figure 6) shows a moderate diagonal, indicating reasonable accuracy across most spectral types. However, there are misclassifications mainly for A, F and G Spectral types, highlighting lack of complexity in the model affecting classification ability for spectral types with overlapping spectral characteristics.

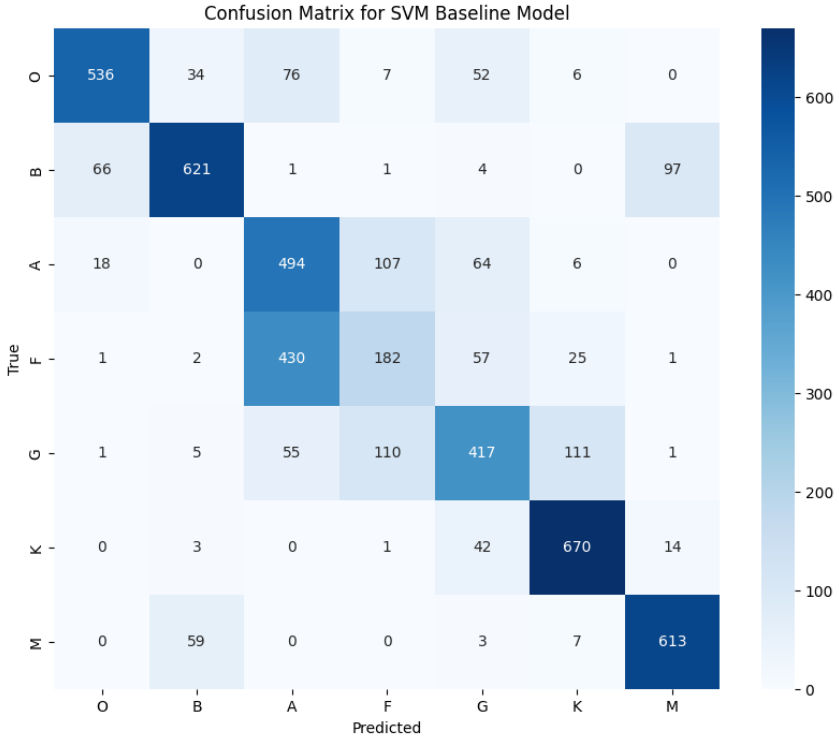

Figure 6: Baseline Model Confusion Matrix

Table 1: Classification Report

	Precision	**Recall**	**F1-score**	**Support**
O	0.86	0.75	0.80	711
B	0.86	0.79	0.82	790
A	0.47	0.72	0.57	689
F	0.45	0.26	0.33	698
G	0.65	0.60	0.62	700
K	0.81	0.92	0.86	730
M	0.84	0.90	0.87	682
Accuracy	-	-	0.71	5000
Macro Avg	0.71	0.70	0.70	5000
Weighted Avg	0.71	0.71	0.70	5000

The ***accuracy*** result for the SVM Baseline model was approximately **70%**, which is much higher than random chance as there are 7 different possible categories. However, it is not an ideal score suggesting the need for a better model. Since the model was reasonable at predicting spectral classifications we use it as the baseline and our goal with the primary model is to improve upon this.

8 Primary Model Quantitative Results

The ***accuracy*** result for your primary model was approximately **84%** out performing our baseline SVM model by **14%**. The model's performance metrics on the 200th epoch: Train error= 0.1644, Train loss= 0.4544, Validation error= 0.1613, Validation loss: 0.4500. The close alignment of training and validation curves indicates effective learning with minimal overfitting.

Figure 7: Error Training Curve

Figure 8: Loss Training Curve

Table 2: Classification Report

	Precision	**Recall**	**F1-score**	**Support**
O	0.91	0.84	0.87	15000
B	0.93	0.91	0.92	15000
A	0.69	0.68	0.68	15000
F	0.66	0.69	0.67	15000
G	0.81	0.84	0.82	15000
K	0.94	0.94	0.94	15000
M	0.96	0.97	0.96	15000
Accuracy	-	-	0.84	105000
Macro Avg	0.84	0.84	0.84	105000
Weighted Avg	0.84	0.84	0.84	105000

Spectral types K and M exhibited the highest F1-scores (0.94 and 0.96, respectively), reflecting their distinct photometric features. In contrast, types A and F showed lower F1-scores (0.68 and 0.67), due to overlapping features, as confirmed in both the confusion matrix and classification report. The balanced dataset ensured fair evaluation, with macro and weighted averages consistently at 0.84.

Quantitative improvements were largely attributed to key preprocessing steps. By using SMOTE to balance the dataset, we ensured equal representation across all spectral types, which significantly improved recall for minority classes such as type O. Additionally, incorporating class weights into the loss function enhanced the model's performance on underperforming classes like A and F, encouraging the model to pay closer attention to these challenging spectral types.

9 Primary Model Qualitative Results

The confusion matrix (Figure 9) highlights a strong diagonal, indicating high accuracy across most spectral types. Misclassifications were more frequent between types A and F, consistent with their overlapping photometric characteristics. Spectral types O, B, K, and M demonstrated minimal misclassification due to their unique features.

Beyond numerical metrics, the HR diagram (Figure 10) validates the model's predictions by closely matching a real HR diagram (Figure 10). Key stellar sequences, such as the main sequence and sub-giant branches, were clearly identified, showcasing the model's effectiveness in capturing astrophysical relationships.

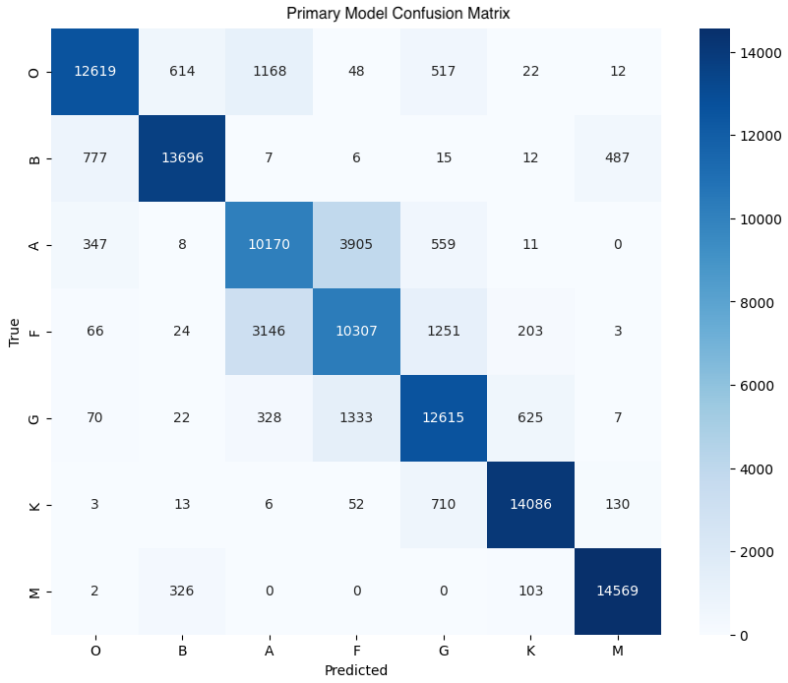

Figure 9: Confusion Matrix of the Baseline Model

10 Evaluation on new data

The model when run on test data that was unseen during training resulted in an accuracy of 84%.

We used the predicted spectral class to get temperatures for each star and created our very own HR diagram based on it, as it is a great way to showcase and validate our model predictions.

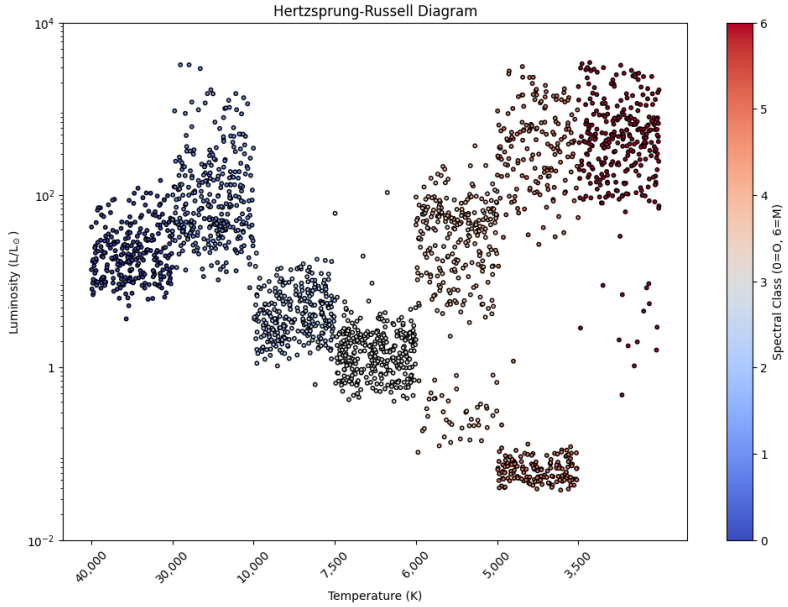

Figure 10: HR diagram created from test data predictions

We can clearly see the similarities between our HR diagram and an official one in Figure 11. There is a line down the middle that resembles the main sequence, which is where most stars tend to be, highlighted in blue. There is a grouping of high luminosity low-temperature stars highlighted in red, which are the sub giants, and an even higher luminosity and lower temperature grouping, which are the giants highlighted in yellow. This shows us that our model was able to predict classes that closely resemble an actual HR diagram, thereby proving that the model is accurate.

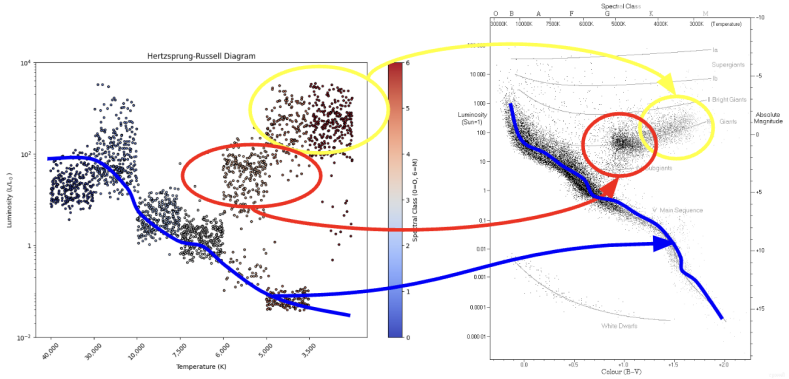

Figure 11: Predicted HR diagram vs Real HR diagram

11 Discussion

The model when run on test data that were unseen during training resulted in an accuracy of 84%. We were able to beat our baseline performance by approximately 14%, which we believe is good performance considering the complexity of the task.

A key insight is the model's ability to handle imbalanced data, particularly the underrepresented O-type stars, which demonstrated consistent predictions despite their rarity. This outcome highlights the success of preprocessing techniques like SMOTE in addressing class imbalances. However, the model showed less accuracy in distinguishing spectral classes A and F, likely due to their overlapping feature spaces. This is evident from confusion matrix analysis and suggests that further refinements may be needed, such as feature engineering or fine-tuning hyperparameters specific to these classes.

Overall, the results demonstrate the model's potential but also highlight areas for improvement, particularly in regions of feature overlap. This project provided valuable insights into the challenges of automating stellar classification and the effectiveness of machine learning in addressing them.

12 Ethical Considerations

A key ethical concern with our model is its dependency on the Gaia dataset, which was used exclusively for training. Different telescopes often use varying instruments, filters, and observational methods, resulting in photometric data that may not align with Gaia's specific measurements (e.g., Gmag, BPmag, RPmag, Lum-flame). Consequently, applying this model to data from other telescopes might lead to inaccurate classifications, reducing its utility and potentially propagating errors in subsequent research.

This limitation risks overgeneralization if users assume the model is universally applicable without considering differences in data acquisition. Stars observed with infrared or ultraviolet telescopes may have characteristics not captured by the Gaia dataset, leading to poor performance or biased results. To mitigate this, users should validate the model on new datasets, retrain it for compatibility when necessary, and provide clear documentation of its training data and limitations to prevent misuse.

13 PROJECT DIFFICULTY/QUALITY

The problem has many challenging aspects which are outlined below**:**

1. **High-dimensional Data:** Stellar classification involves photometric magnitudes and luminosity, which are subtle and interdependent. The relationships between features and spectral classes are nonlinear and complex, making the classification task challenging.

2. **Imbalanced Data:** Spectral classes such as O-type stars are underrepresented in the dataset. Handling this imbalance during training, without overfitting to more frequent classes like M-type stars, required careful preprocessing, including using new techniques like SMOTE for oversampling that we did not learn in labs.

3. **Astronomical Nuances:** The spectral classifications inherently include overlapping characteristics. For instance, stars near class boundaries (e.g. between F and A as we said earlier) can be difficult to differentiate.

4. **Training Challenges and Advanced Techniques not learned in labs:** To enhance model performance, we implemented weighted loss functions to address class imbalances, particularly for spectral types A and F, ensuring the model penalized errors in these underrepresented classes more effectively. We also utilized advanced techniques such as CUDA for faster computations, LeakyReLU for improved gradient flow in negative regions, and a step-based learning rate scheduler to fine-tune the optimization process. These adjustments significantly improved training stability and accuracy.

Considering all these difficulties, we do believe a test accuracy of 84% is reasonably good, especially since we beat our baseline SVM model by 14%. In addition, the HR diagram generated by our model accurately reflects stellar trends such as the main sequence and giants, further validating the effectiveness of the model in classifying stars.

Conclusion: The project was inherently challenging due to the domain complexity, imbalanced classes, and indirect feature-space mapping. Despite these difficulties, our model performed better than expected, demonstrating its robustness and our ability to apply advanced machine learning techniques effectively.

14 REFERENCES

REFERENCES

[1] W. W. Morgan, H. A. Abt, and J. W. Tapscott, "Revised MK spectral atlas for stars earlier," https://ned.ipac.caltech.edu/level5/March02/Morgan/paper.pdf, accessed Oct. 5, 2024.

[2] R. O. Gray and C. J. Corbally, Stellar Spectral Classification, Dec. 2009.doi:10.1515/9781400833368.

[3] R. Andrae, D.-W. Kim, and C. A. Bailer-Jones, "Assessment of stochastic and deterministic models of 6304 quasar lightcurves from SDSS Stripe 82," Astronomy & Astrophysics, vol. 554, Jun. 2013. doi:10.1051/0004-6361/201321335.

[4] S. Fabbro et al., "An application of deep learning in the analysis of Stellar Spectra," OUP Academic, https://academic.oup.com/mnras/article/475/3/2978/4775133, accessed Oct. 4, 2024.

[5] N. Mowlavi et al., "Large-amplitude variables in Gaia data release 2," Astronomy & Astrophysics, https://www.aanda.org/articles/aa/full_html/2021/04/aa39450-20/aa39450-20.html, accessed Oct. 4, 2024.