Half a Million Binary Stars Identified from the Low-resolution Spectra of LAMOST

Yingjie Jing[1], Tian-Xiang Mao[1], Jie Wang[1,2,3], Chao Liu[4,3,2], and Xiaodian Chen[5,3,2]

Reporter: Xu Zhang

2025.6.20

Outline

1. Introduction
2. Data and Methods
3. Results
4. Conclusions and Discussions

1. Introduction

Binary Stars: Ubiquitous and Fundamental

- Binary and multiple star systems comprise roughly half of all stars
- Essential for precise measurements of mass, radius, and luminosity (Torres & Andersen 2010; Eker et al. 2018)
- Crucial for testing theories of stellar formation and evolution, galactic archeology, gravitational waves (Duquennoy & Mayor 1991; Raghavan et al. 2010; Moe & Di Stefano 2017)

Traditional Detection Methods:

- Radial velocity variations (Pryor et al. 1988; Cote et al. 1994)
- Brightness variations/light curves (Yan & Mateo 1994; Albrow et al. 2001)
- Color-magnitude diagram analysis (Sollima et al. 2010; Li et al. 2013)

Limitations: Require multiple observations, effective mainly for bright, close binaries

1. Introduction

Large Spectroscopic Surveys: New Opportunities

- APOGEE (Holtzman et al. 2015), RAVE (Steinmetz et al. 2006), LAMOST (Cui et al. 2012)
- Previous catalogs: hundreds (RAVE) to thousands (APOGEE) of binaries
- LAMOST studies identified hundreds of thousands (Qian et al. 2019; Liu et al. 2024)
- Gaia: over one million binaries (El-Badry et al. 2021)

New Approach: Convolutional Neural Networks(CNNs)

- CNNs excel at complex pattern recognition (LeCun et al. 1989; Krizhevsky et al. 2012)
- Data-driven approach eliminates manual feature engineering
- Successfully applied to astronomical problems (Ting et al. 2018; Davies et al. 2019)

2. Data and Methods

2.1. Data Set

Training Sample from C. Liu (2019):

- Solar neighborhood sample based on H-R diagram positions
- Binary sequence located above single main sequence

Selection criteria:

- Single stars: $-0.25 < \Delta M_G < 0.25$ (black points)
- Intermediate-mass-ratio binaries: $-0.5 < \Delta M_G < -0.25$ (blue points, mass ratio ~0.71-0.93)
- Excluded high-mass-ratio binaries: $\Delta M_G < -0.5$ (red points)

Final Training Sample: 68,299 single stars and 3,818 binary stars

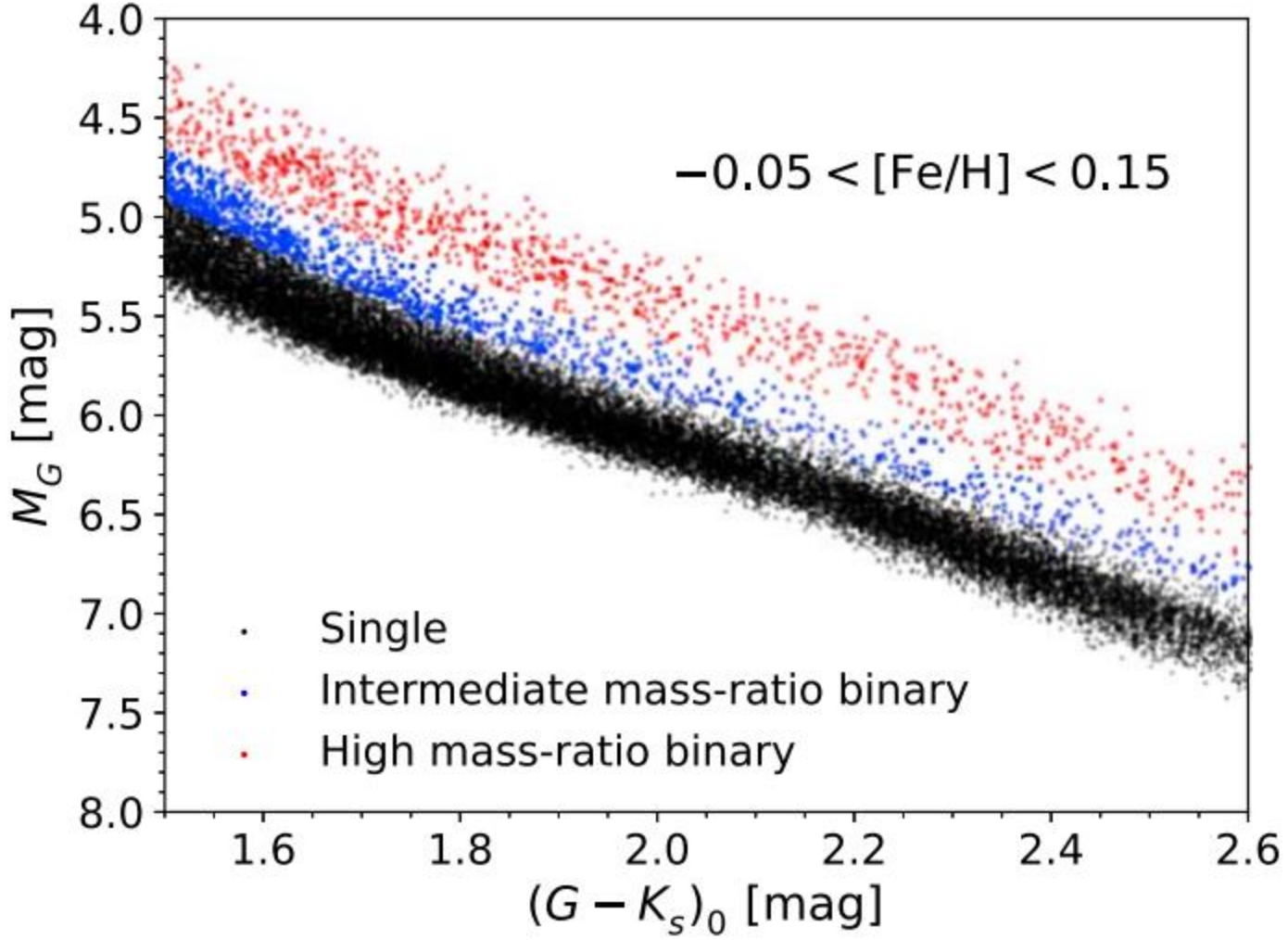

Figure 1. Color–magnitude diagram for stars with $-0.05 < [Fe/H] < 0.15$ from the training sample. The black, blue, and red points represent identified single main-sequence, intermediate-, and high-mass-ratio binary stars, respectively. The single stars and intermediate-mass-ratio binaries are selected as a training sample. The G-band absolute magnitude (M_G) is plotted against the dereddened color index $(G - K_s)_0$.

2. Data and Methods

2.1. Data Set

Initial Stellar Sample Construction:

- Cross-matching: LAMOST DR10 LRS A, F, G and K stars catalog+ Gaia DR3 + 2MASS (Skrutskie et al. 2006)
- ~7 million spectra initially
- LAMOST low-resolution spectra: 3700-9000 Å, R ≈ 1800 (Cui et al. 2012; Luo et al. 2012)

Selection Criteria (from C. Liu 2019):

1. $3800 < T_{eff} < 6500$ K
2. $\log g > 4$
3. $1.5 < (G - Ks)_0 < 2.6$
4. SNR > 20 in g band
5. $\varpi > 3$ mas (excluded in this work)
6. $M_G > 4$ mag (excluded in this work)

Final Sample: 1,258,912 spectra from 971,805 stars

2. Data and Methods

2.2. Spectral Data Preprocessing

Data Preprocessing Steps:

1. Data cleaning: Remove spectra with excessive masked data points

2. Interpolation: Onto new wavelength grid

3. Normalization: Divide by smoothed flux $f_s(\lambda)$ (A. Y. Q. Ho et al. 2017)

Smoothed flux definition:

$$f_s(\lambda) = \frac{\sum_i (f_i w_i(\lambda))}{\sum_i (w_i(\lambda))},$$

where $w_i(\lambda)$ is Gaussian function $w_i(\lambda) = e^{-\frac{(\lambda-\lambda_i)^2}{\sigma^2}}$. with $\sigma = 35$ nm.

Data Splitting: 80% training, 10% validation, 10% test

Class Imbalance: Binary:Single ≈ 1:18, solved by oversampling (Buda et al. 2017)

2. Data and Methods

2.3. Deep Learning Model

CNN Architecture:

- Basic residual learning block (He et al. 2015)
- Two fully connected layers
- 1D convolution for spectral data
- **Convolutional layer definition:**

$$\boldsymbol{x}_n^l = a\left(\sum_k \boldsymbol{W}_n^l \otimes \boldsymbol{x}_{n-1}^k + \boldsymbol{b}_n^l\right).$$

- **Overfitting prevention:** Dropout layers (rate 0.5), early stopping (Srivastava et al. 2014)

Training Parameters:

- Cross-entropy loss function
- Learning rate: 1×10^{-4}
- **Output:** Probability $p_b \in [0,1]$ of being binary

3. Results

3.1. Test Set Performance

ROC Curve Analysis:

- **Area under ROC curve: 0.949**

(significantly better than random 0.5)

- 95.5% of single stars: $p_b < 0.2$
- 73% of binary stars: $p_b > 0.8$

Precision Analysis:

$$\text{Precision} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}} \times N_{\text{b}}/N_{\text{s}}/r}$$

- **Precision = 0.79** ($p_{th} = 0.5$)
- **Precision = 0.89** ($p_{th} = 0.8$)
- Assuming binary-to-single ratio 1:2

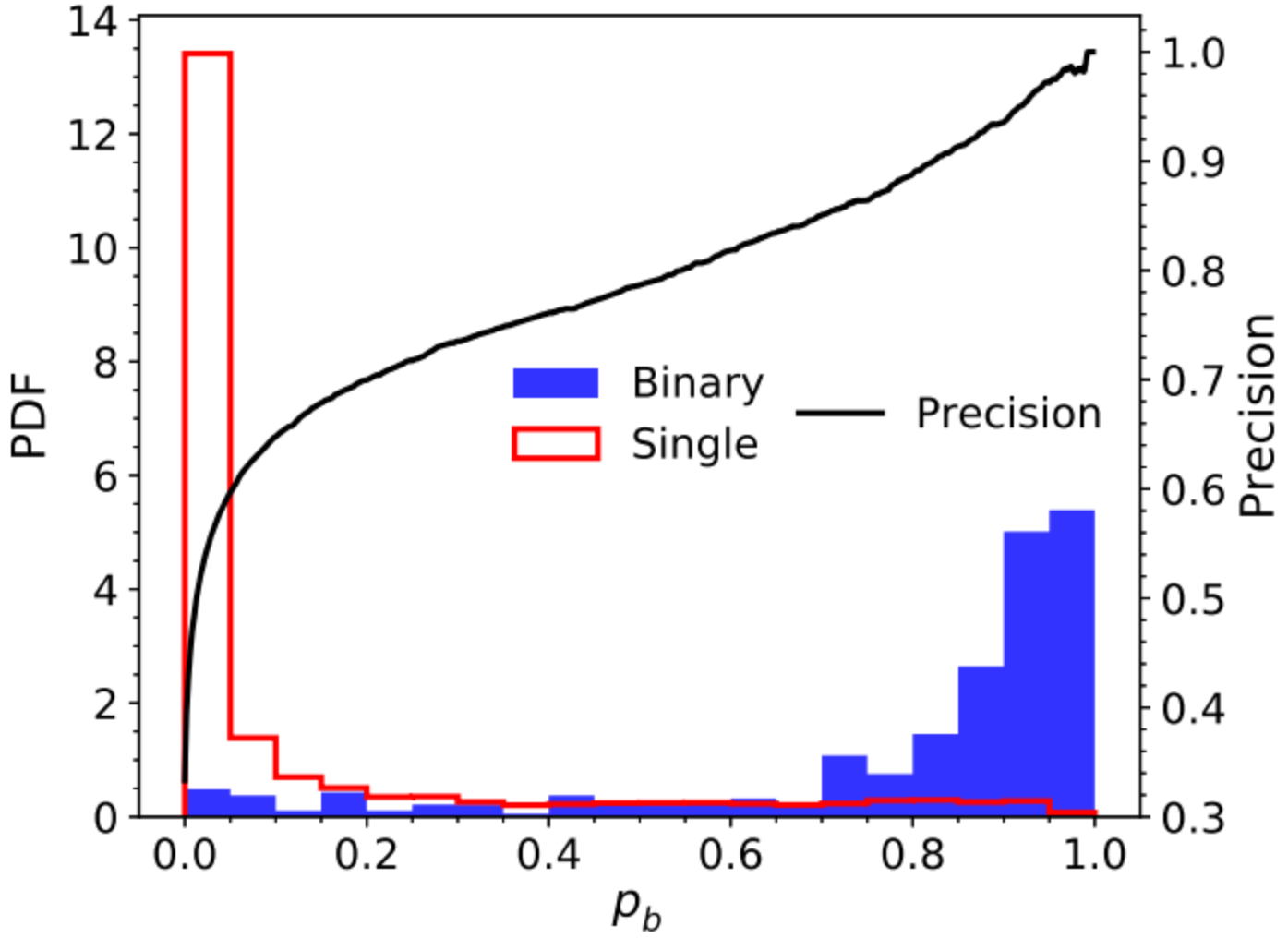

Figure 3. Probability density function (PDF) of the predicted probabilities (p_b) for binary (blue region) and single (red line) stars in the test set. The black line shows the precision as a function of the adopted probability cutoff threshold, assuming a binary-to-single-star ratio of 1:2.

3. Results

3.1. Test Set Performance

Mass Ratio Sensitivity:

Highest binary fraction for mass ratios ~0.71-0.93 (training range)

Decreased sensitivity for:

- High mass ratios (q → 1): excluded from training due to spectral similarity
- Low mass ratios: limited by training data ΔM_G cuts

Overall trend influenced by training data selection

Key Finding: Network performs best for intermediate-mass-ratio binaries as expected from training design

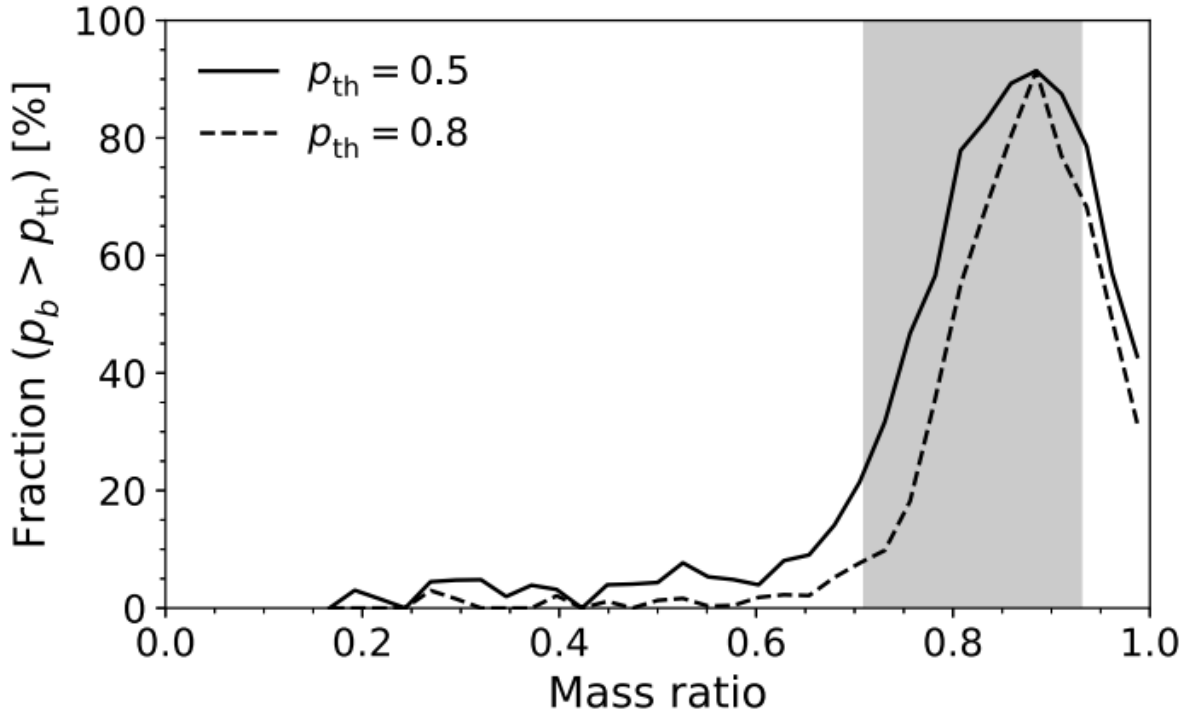

Figure 4. The fraction of binary stars as a function of mass ratio for different p_{th} values. The shaded area denotes the regions of intermediate-mass-ratio binaries (corresponding to mass ratios between approximately 0.71 and 0.93), which is included in the training sample.

3. Results

3.2. Comparison with Other Methods

Eclipsing Binaries Validation:

• Sample: 535 EA + 2724 EW eclipsing binaries from Chen et al. (2020)

• Zwicky Transient Facility (ZTF): 350,000 EBs detected

•Detection rates (p_{th} = 0.5):

- **EA binaries: 96.45%**
- **EW binaries: 96.40%**

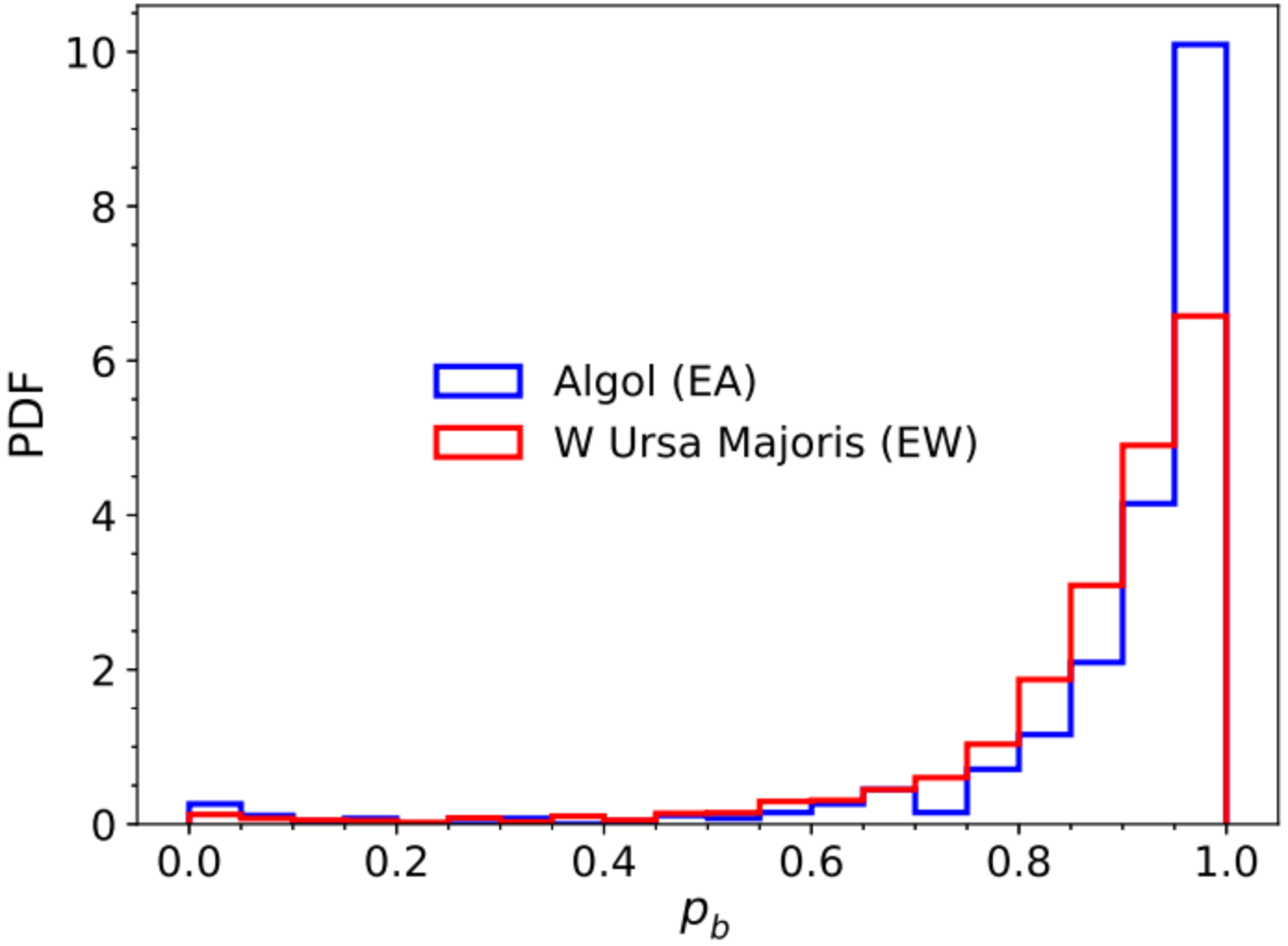

Figure 5. Probability density function (PDF) of the predicted binary probabilities (p_b) for eclipsing-binary stars. The blue solid line corresponds to Algol-type (EA) binaries, while the red solid line corresponds to W Ursa Majoris–type (EW) binaries, demonstrating a high detection rate of 96.45% for EA and 96.40% for EW binaries.

3. Results

3.2. Comparison with Other Methods

Radial Velocity Validation:

- Potential binaries: $\Delta_{rv,max} > 25$ km s^{-1}
- Likely singles: $\Delta_{rv,max} < 5$ km s^{-1} ($\geq$4 observations)
- **Agreement rates ($p_{th} = 0.5$):**
  - **Binary stars: 92.8%**
  - **Single stars: 68.8%** (lower due to undetected long-period binaries)

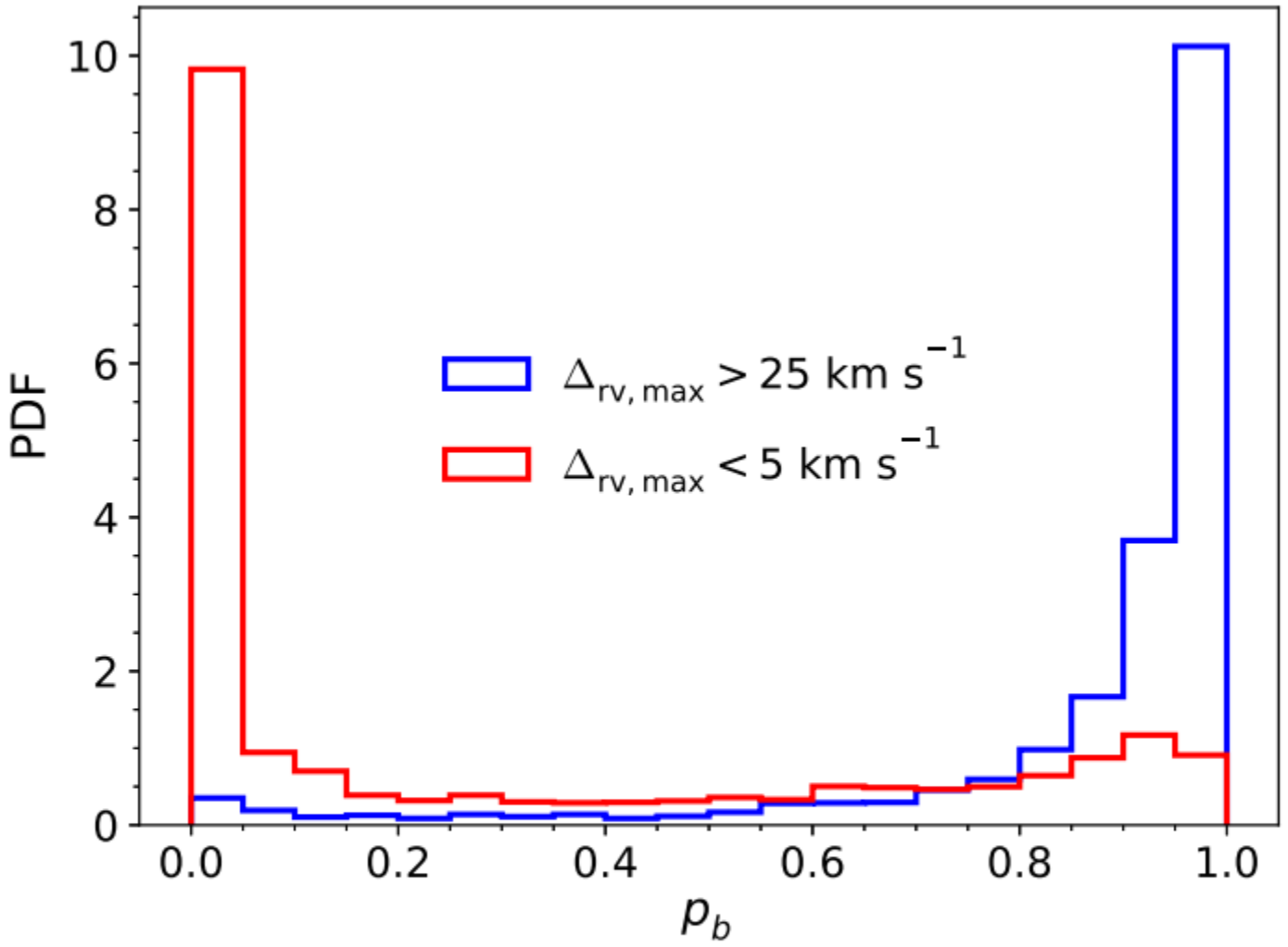

Figure 6. Probability density function (PDF) of the predicted binary probabilities (p_b) for stars observed multiple times in the LAMOST survey. The blue line indicates potential binary stars (with $\Delta_{rv,max} > 25$ km s^{-1}), while the red line indicates likely single stars (with $\Delta_{rv,max} < 5$ km s^{-1}).

3. Results

3.3. Main-sequence Binary-star Catalog

Catalog Statistics:

468,634 binary stars ($p_b > 0.5$)

323,909 stars with $p_b > 0.8$ (high confidence)

Threshold $p_b > 0.5$: compromise between accuracy and purity

Comparison with Previous Studies:

Qian et al. (2019): 256,000 spectroscopic binaries (LAMOST, multiple observations)

Jack (2019): 34,691 spectroscopic binaries (Gaia DR2)

Birko et al. (2019): 27,716 single-lined binaries (RAVE + Gaia DR2)

Price-Whelan et al. (2018): 4,898 SB1 (APOGEE red giants)

This catalog surpasses previous spectroscopic binary studies in size

3. Results

3.3. Main-sequence Binary-star Catalog

Catalog Characteristics:

Probability Distribution:

Most spectra have $p_b \approx 0$ or $p_b \approx 1$ (clear classification)

For multiple observations: p_b = maximum value

Higher $p_{th} \rightarrow$ fewer binaries but higher precision

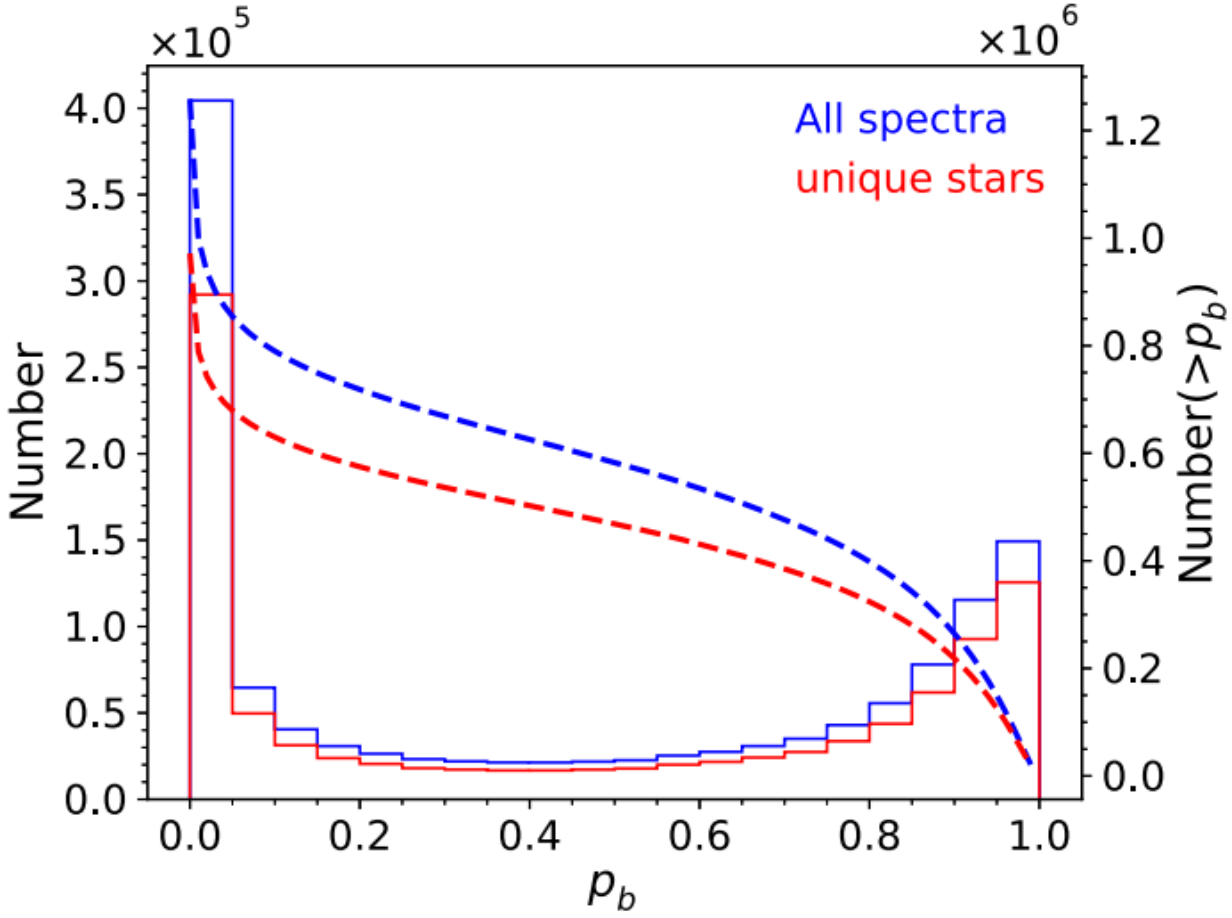

Figure 7. The distribution of p_b provided by the network for all spectra (blue) and individual stars (red) within the main-sequence star sample. The solid curves represent the number distribution of p_b, while the dashed lines indicate the count of spectra (or individual stars) with as a function of the p_{th}. For individual stars, p_b corresponds to the highest value in cases of multiple observations.

3. Results

3.3. Main-sequence Binary-star Catalog

Catalog Characteristics:

Spatial and Distance Distribution:

Distance range: up to ~19 kpc

Median distance: ~0.7 kpc for binary stars

115 binary stars beyond 10 kpc from the Sun

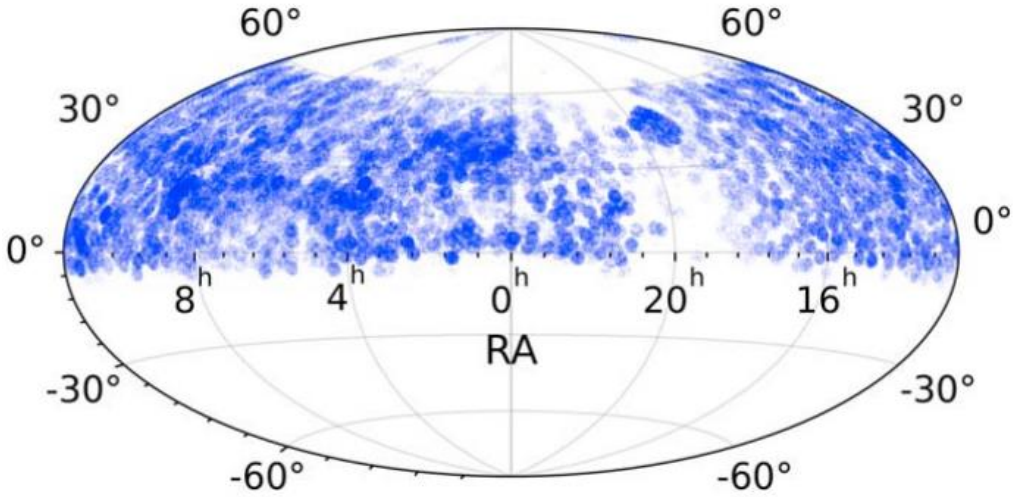

Figure 8. Distribution of binary stars in the R.A. and decl. plane.

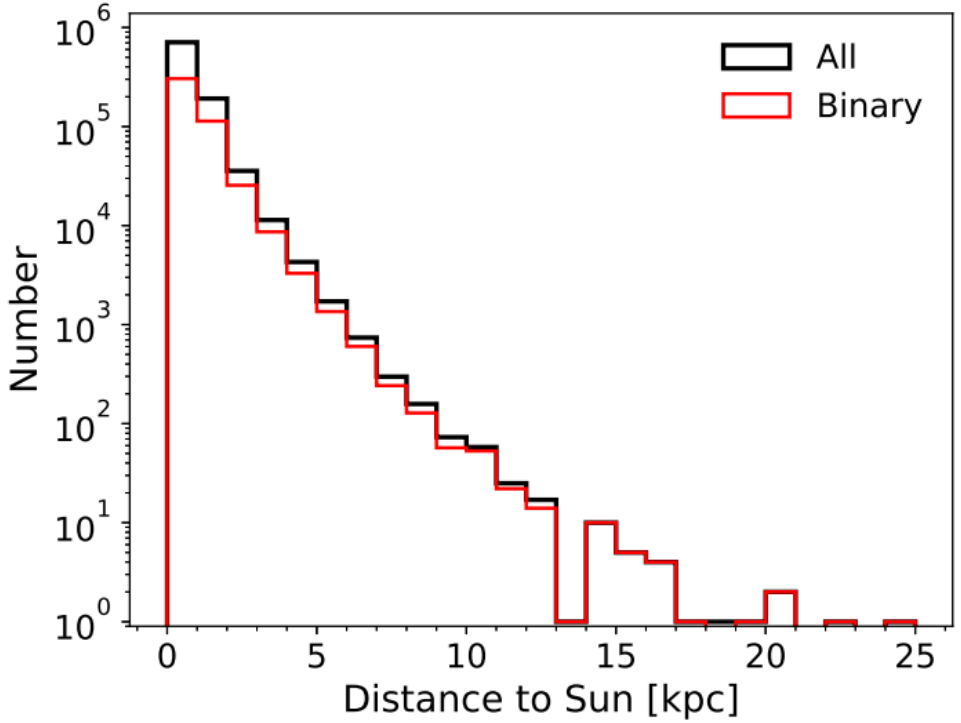

Figure 9. Distribution of distances to the Sun for all stars (black) and binary stars (red). The histogram reveals that binary stars are found at a range of distances, with 115 binary stars situated beyond 10 kpc.

3. Results

3.3. Main-sequence Binary-star Catalog

Catalog Characteristics:

Color-Magnitude Diagram Validation:

Predicted binaries appear above single-star main sequence (as expected)

Some $p_b < 0.5$ stars above binary sequence: likely high-mass-ratio binaries

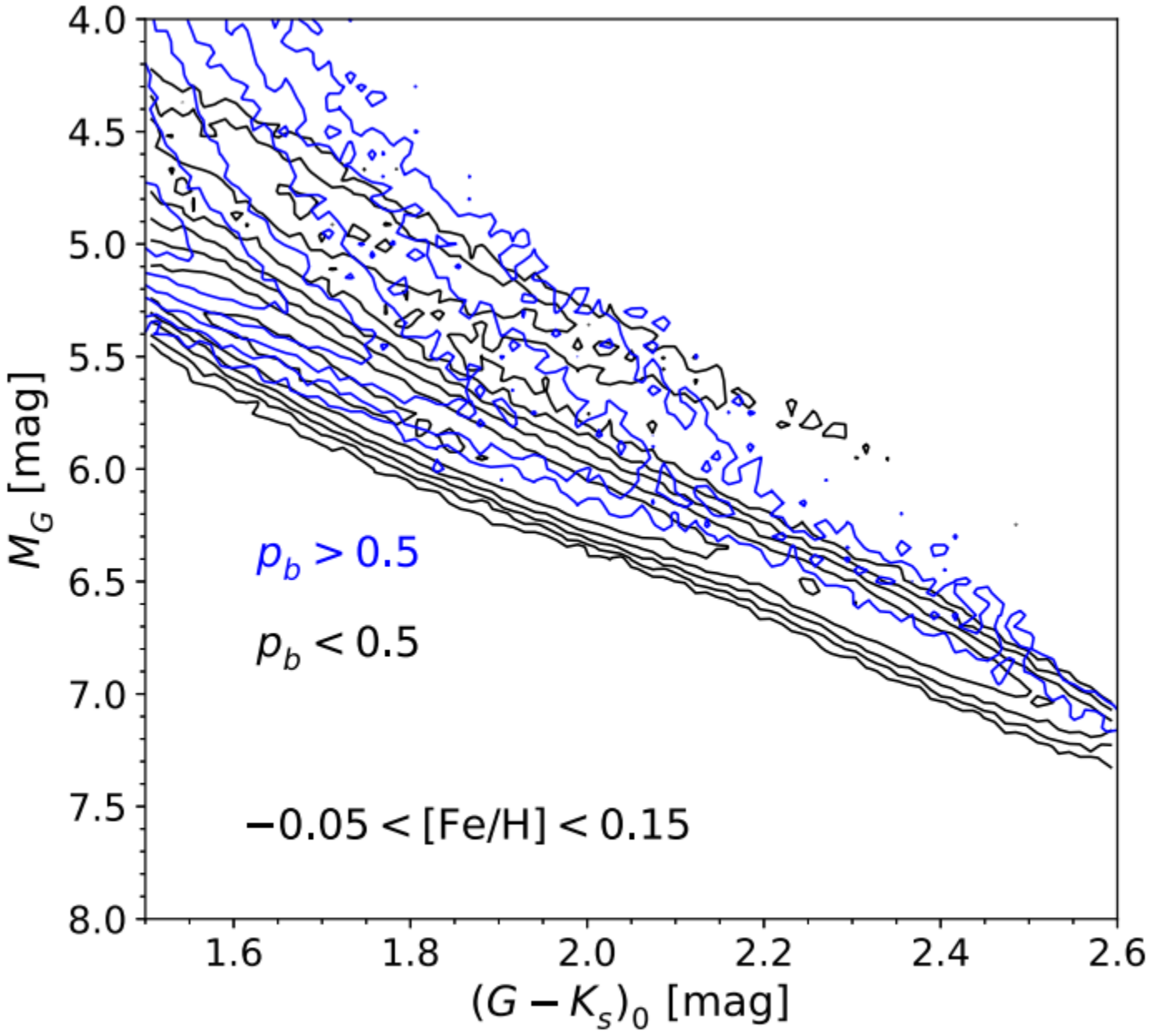

Figure 10. Number density contour plot in the color–magnitude diagram of stars with $-0.05 < [Fe/H] < 0.15$. Stars with $p_b < 0.5$ are shown in black contours, while those with $p_b > 0.5$ are highlighted in blue.

3. Results

3.4. Planets in Binary-star Systems

Exoplanet Host Discovery:

Cross-matched with NASA Exoplanet Archive (accessed 2024 Sep 24)

128 binary systems hosting confirmed exoplanets

114 previously unidentified as binaries in the archive

Scientific Significance:

Demonstrates existence of planets in binary systems (Desidera & Barbieri 2007; Mugrauer & Neuhäuser 2009)

Important for planet formation studies in dynamically complex environments (Thebault & Haghighipour 2015)

Diverse binary configurations and planetary architectures

4. Conclusions and Discussions

Key Achievements:

1. Novel CNN approach for binary identification using single-epoch spectra
2. High performance: ROC area 0.949, ~96% detection rate for eclipsing binaries
3. Largest spectroscopic binary catalog: 468,634 binary stars
4. 128 binary systems with exoplanets discovered

Method Advantages:

- Single-epoch spectra vs. traditional multi-epoch radial velocity methods
- Particularly valuable for large spectroscopic surveys like LAMOST
- Less affected by distance compared to photometric methods

Limitations and Future Work:

Current limitations:

Training focused on main-sequence stars only
Potential biases in training data selection
Difficulty detecting highest-mass-ratio binaries

Future directions:

Expand training dataset to include wider range of stellar types
Explore alternative training data selection methods
Mitigate potential biases in training sample

Thanks!