Mining double-line spectroscopic candidates in the LAMOST medium-resolution spectroscopic survey using human-AI hybrid method

Shan-shan Li [1,2] Chun-qian Li [3,2] Chang-hua Li,[1] Dong-wei Fan [1] Yun-fei Xu [1] Lin-ying Mi,[1,2]
Chen-zhou Cui [1,2] and Jian-rong Shi [3,2]

arXiv:2411.14714v1
Reporter: Xu Zhang

- **Introduction**
- **DATA SELECTION**
- **THE METHOD**
- **Result**
- **Discussion**
- **CONCLUSION AND PROSPECTS**

Introduction

Introduction

•Multiple star systems, also called multi-systems, including binary systems play an essential role in modern astrophysics.

•Approximately half of the stars in the Milky Way reside in multiple star systems. (Duchene & Kraus 2013; Raghavan et al. 2010)

•Binary star systems serve as the only known direct method to determine accurate stellar masses.

•Multiple star systems provide insights into:

- Processes of stellar formation and evolution
- Material exchange and interactions
- Evolutionary paths and final outcomes

Spectroscopic Binaries (SBs)

Due to the Doppler shift resulting from different radial velocities (RVs):

SB1s:

•Only the spectrum of one star can be clearly observed

•Identified by variations in RV

SB2s:

•Both stars have similar spectral types but significant RV differences

•Can be observed as double-lined spectroscopic binaries

•Identified by double peaks in CCF

Recent surveys have made systematic SB detection possible:

- **APOGEE Survey**: Over 7000 SB2s identified from DR16 (Kounkel et al. 2021)
- **Gaia-ESO Survey**: 641 SB1s, 342 SB2s, and 11 SB3s identified (Merle et al. 2017, 2020)
- **GALAH Survey**: Large contribution to SB studies (Traven et al. 2020)

LAMOST Survey

- Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST)
  - Special reflecting Schmidt telescope
  - Enables survey with up to 4000 optical fibers
- LAMOST Medium-Resolution Spectroscopic Survey (LAMOST-MRS)
  - Began in September 2018
  - Resolving power of R ~ 7500
  - Capable of measuring RVs with high precision
- Previous Works:
  - Li et al. (2021): 3133 SB2 and 132 SB3 candidates using DR7
  - Zhang et al. (2022): 2198 SB2 candidates using DR8
  - Kovalev et al. (2022): 2460 SB2 candidates

DATA SELECTION

Data and Preprocessing

- LAMOST-MRS DR9 (released in March 2022):
  - 29,920,588 spectra in total
  - All data processed by LAMOST Stellar Parameter Pipeline (LASP)
  - Includes time-domain and non-time-domain surveys
- Each exposure in LAMOST-MRS yields two spectra:
  - Blue arm (4950 ~ 5350 Å)
  - Red arm (6300 ~ 6800 Å)
- Blue arm spectra selected due to more absorption lines

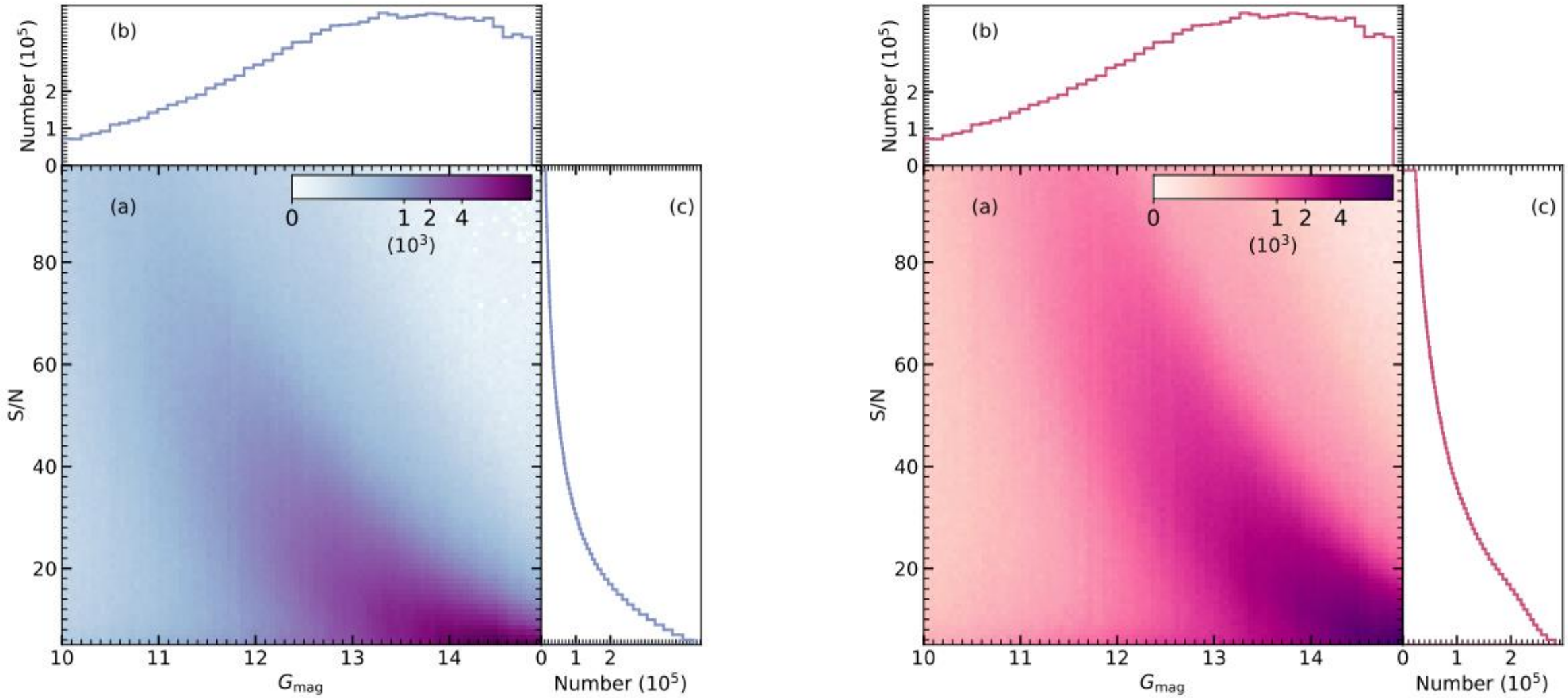

Figure 1. The left panel displays the spectra from the blue arm of the LAMOST-MRS DR9 data, while the right panel presents the spectra from the red arm. In each panel, the parts labeled as a, b, and c represent the distribution of S/N versus G magnitude, the distribution of G magnitude, and the distribution of S/N, respectively.

Table 1. The criteria we used for screening LAMOST-MRS DR9 data.

Field	Criteria	Notes
band	B	Only blue arm spectrum
coadd	0	No combined spectrum
fibermask	0	Delete spectra from bad fiber
bad_b	0	Delete bad blue arm spectra
S/N	≥ 5	

They have obtained a final sample of 6,565,721 spectra from 930,783 stars.

THE METHOD

The Method

•Two main steps in their human-AI hybrid process:

1.Traditional CCF Technique

1. Calculate RVs using observed and template spectra
2. Generate categorized list of spectra
3. Produce corresponding CCF data

2.Machine Learning

1. Apply multiple ML classifiers
2. Use ensemble learning strategy
3. Identify SB2 and SB3 systems

Cross-Correlation Function

- CCF Calculation:
  - Based on observed and template spectra
  - Range: -500 to +500 km/s
  - Step size: 1 km/s

$$CCF(v) = \sum_{i=1}^{n} \left(\frac{O_i - \overline{O}}{\sigma_O}\right)\left(\frac{T_{i,v} - \overline{T}}{\sigma_T}\right)$$

- Template: generated by spectral synthesis program SPECTRUM (Gray & Corbally 1994) and ATLAS stellar atmospheric models (Castelli & Kurucz 2003)
- Three Template Spectra:
- Hot dwarfs: Teff = 8000 K, log g = 4.0 dex, [Fe/H] = 0.0 dex
- Cool dwarfs: Teff = 5000 K, log g = 4.0 dex, [Fe/H] = 0.0 dex
- Cool giants: Teff = 5000 K, log g = 2.0 dex, [Fe/H] = 0.0 dex
- After calculating the CCFs, the highest of the three CCFs is chosen for the next step in the multi-line spectral detection process.

RV Component Detection

- Method from Merle et al. (2017):
  - Calculate first three derivatives of CCF
  - Locate peak positions including blended peaks
  - Use third derivative zero-crossing for RV measurement
- Gaussian Smoothing: Using scipy.ndimage.gaussian_filter1d
  - Initial $\sigma = 13$ km/s
  - Increase by 1 km/s steps
  - Until matching valley numbers or σ reaches 100 km/s
- Selection Criteria:
  - CCF values > 60%
  - Second derivative < 40%

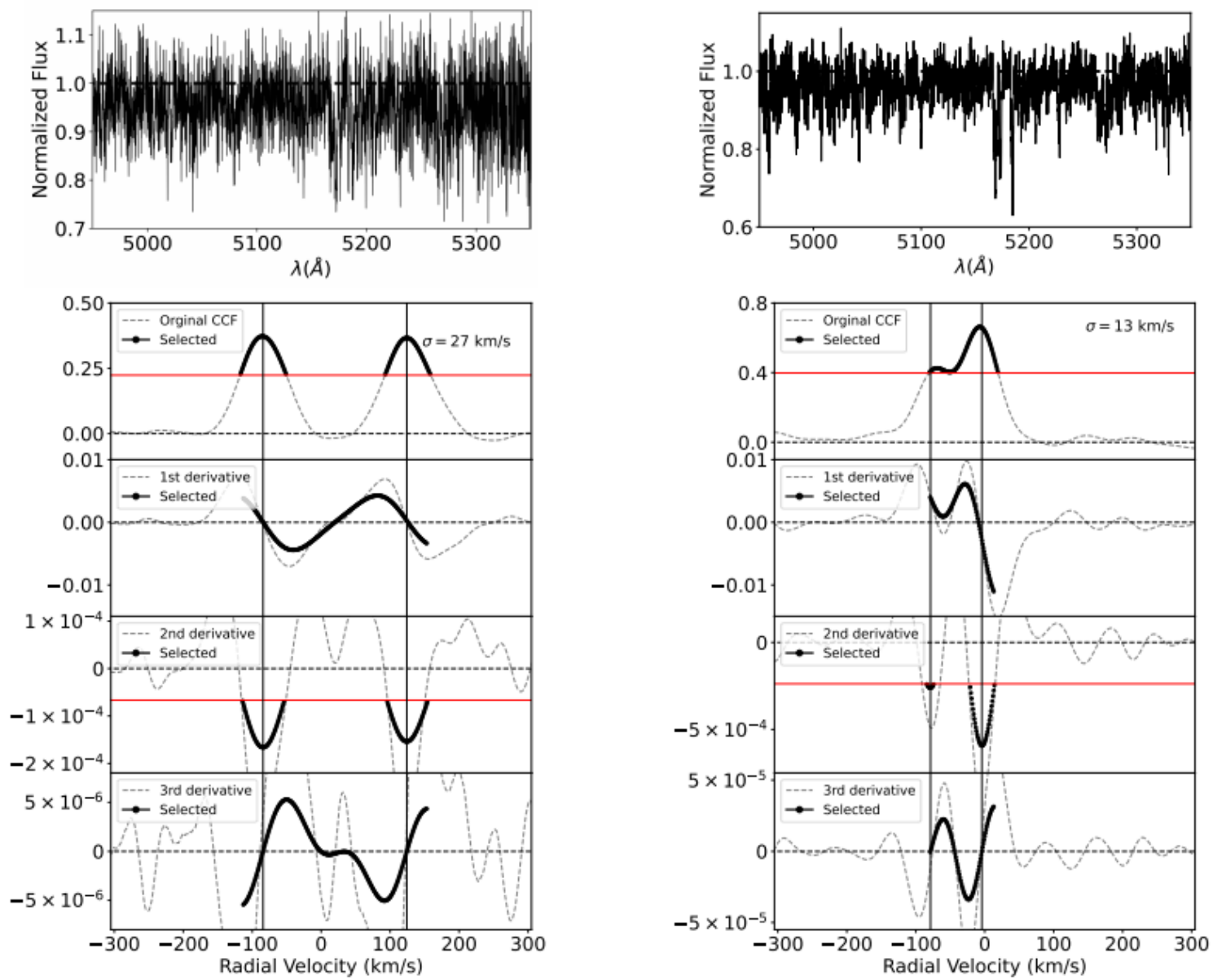

Figure 2. The normalized spectra, CCFs, and derivatives of two SB2 candidates. In the left panel, the final σ of the system is 27 km/s. In the right panel, the CCF exhibits significant peak blending; the first derivative cannot distinguish the peaks well, but they can be identified using the third derivative. Black solid lines are used to draw the selected range of smoothed CCFs and derivatives, while gray dashed lines illustrate the original CCFs. Red horizontal lines indicate the thresholds (above 60% for CCF values and below 40% for the second derivative of the CCF), and black vertical lines mark the RVs.

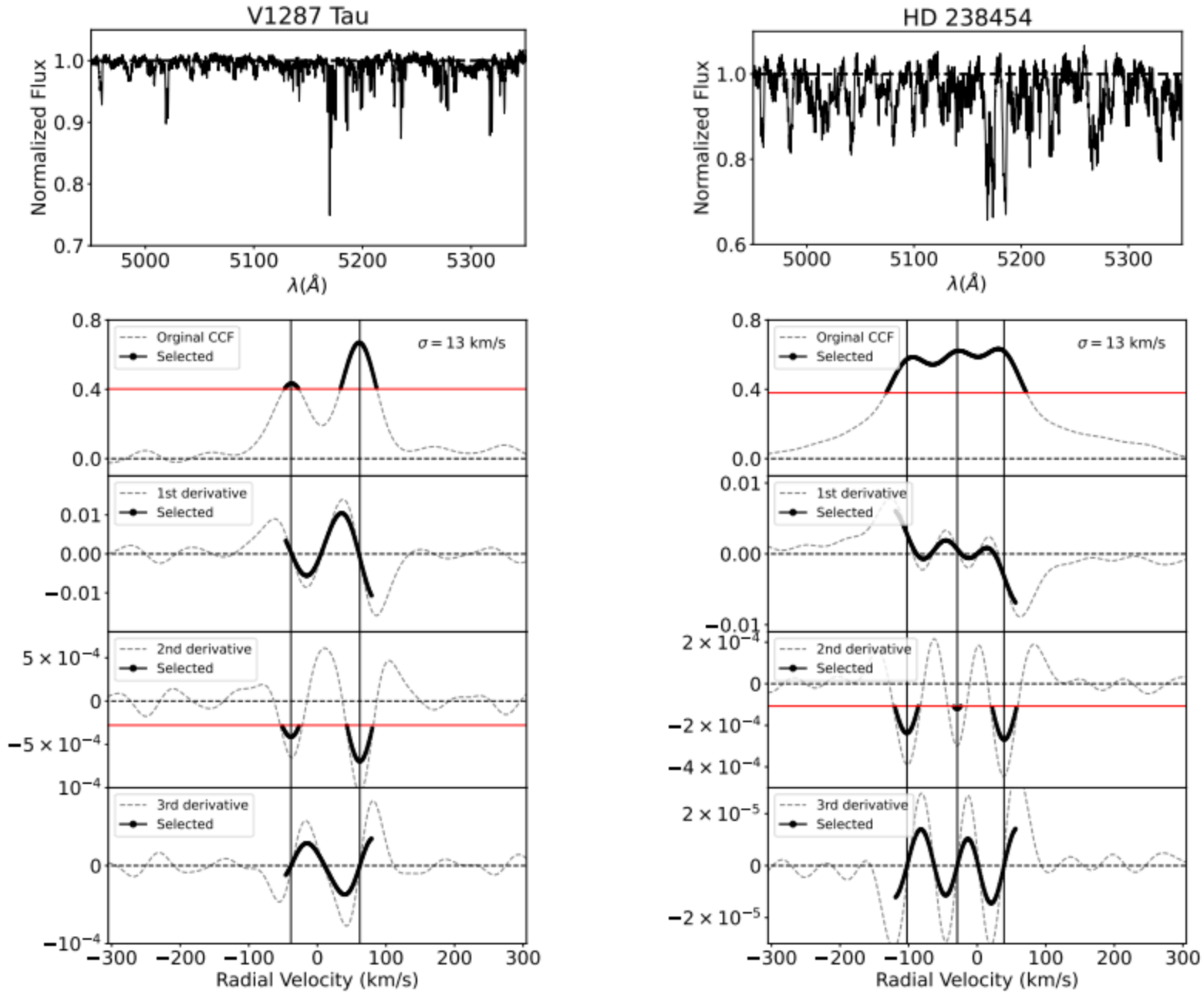

Figure 3. The normalized spectra, CCFs and derivatives of SB2 (V1287 Tau) and SB3 candidates (HD 238454) selected from LAMOST-MRS. The Black solid lines, gray dashed lines, red horizontal lines, and black vertical lines are used as in Figure 2.

Machine learning optimization

•They utilized deep neural networks (DNNs), a type of multi-layer supervised learning model.

Training Dataset Preparation

•Four Categories:

- L0: CCFs without significant peaks
- L1: Single-line spectra CCFs
- L2: Double-line spectra CCFs
- L3: Triple-line spectra CCFs

•Synthetic Spectra Parameters:

- Teff: 3500-8000 K
- log g: 0.0-5.0 dex
- [Fe/H]: -4.0-0.5 dex
- RV difference: 60-250 km/s

•Total Training Samples:

- 1500 samples each from L0, L1, L2, L3
- Total 6000 training sample

DNN Architecture:

•Three fully connected layers:

- First layer: 600 neurons with ReLU activation
- Second layer: 300 neurons with ReLU activation
- Output layer: n neurons with softmax activation (n = number of classes)

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, \quad i = 1, 2, \cdots, n$$

Training Process:

•Framework: Keras with TensorFlow(Abadi et al. 2016)

•Optimizer: Adam (Adaptive Moment Estimation)

•Loss function: Sparse Categorical Cross Entropy $LOSS = -\frac{1}{N}\sum_{i=1}^{N} \log(P_{y_j})$

•Output: Normalized probabilities between 0 and 1

Classification Criteria

- Four Classifiers (C1-C4):
  - C1: Direct four-category classification
  - C2: Two-step (3 categories + binary)
  - C3: Three categories (L0, L1, L2)
  - C4: Three categories (L0, L1, L3)
- Classification Thresholds:
  - SB2: L2 probability > 95% in C1, C2, C3
  - SB3: L3 probability > 99% in C1, C2, C4
- Final Results:
  - 27,233 CCFs labeled as L2
  - 11,904 CCFs labeled as L3

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ \ score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

The evaluation results are very promising, as all values are greater than 99%. This indicates that their classifiers perform very well on the simulated data.

Table 2. The classification results of all calculated CCFs. The last column represents the number of spectra classified as SB2 or SB3 by all machine learning classifiers. For SB2, normalized probabilities (scores) must be greater than 95%. For SB3, the selection criterion is set at 99%.

RV calculation classification	Classifier	L0	L1	L2	L3	Selected results
SB2 (118,274)	C1	52,661	181	37,913	27,519	27,233 ($P > 95\%$)
	C2	44,219	1157	38,290	34,608	
	C3	57,436	208	60,630	-	
SB3 (49,847)	C1	22,445	6	549	20,519	11,904 ($P > 99\%$)
	C2	20,758	2	676	22,083	
	C4	23,365	44	-	20,110	

RESULT

Initial Detection Results

- Analysis of 6,565,721 blue-arm spectra from LAMOST-MRS DR9
- Traditional CCF technique identified:
  - 118,274 double-line spectra
  - 43,519 triple-line spectra
- After machine learning classification:
  - 27,233 double-line spectra
  - 11,904 triple-line spectra
- Final confirmation after visual inspection:
  - 27,164 double-line spectra
  - 3,124 triple-line spectra

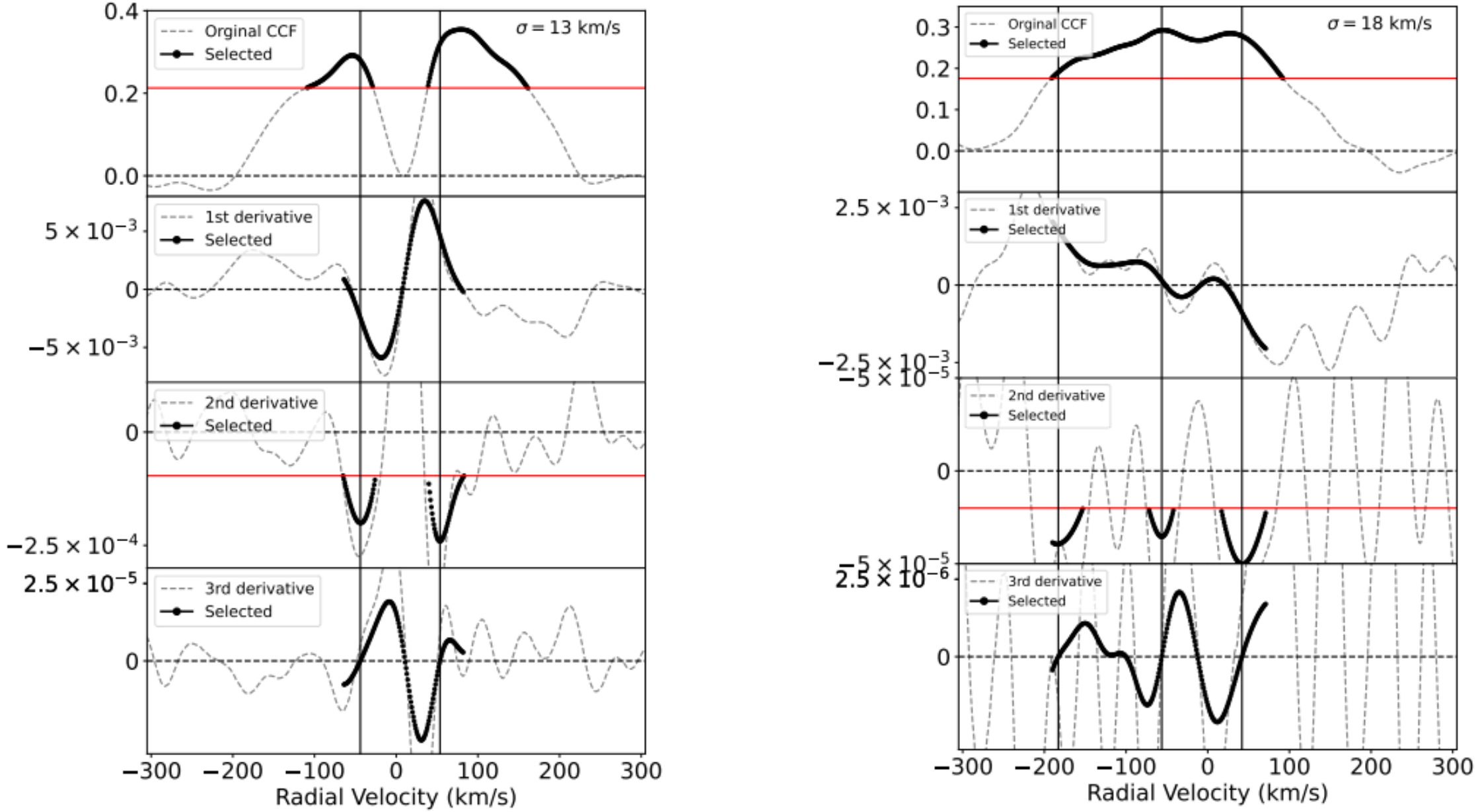

Figure 7. Two examples of cases rejected by visual inspection, including the CCF and its derivatives. The left shows a double-line case identified by the CCF technique and ensemble learning but rejected based on the visual inspection criterion, while the right shows a similar triple-line case.

Cross-matching Results

•Matched with existing catalogs: the Kepler Eclipsing Binary Stars (KEBC, Kirk et al. 2016), the TESS Eclipsing Binary stars (TESS-EBs, Prsa et al. 2022), SB in the APOGEE DR16 and DR17 Data (Kounkel et al. 2021), Gaia DR3 Non-single stars (Gaia Collaboration 2022), SB candidates from Gaia-ESO Survey (Merle et al. 2017), etc:

- 690 common binary candidates
- 151 common triple-star candidates

•Matched with the candidates obtained by Li et al. (2021) using LAMOST-MRS DR7 data:

- 1637 SB2 and 58 SB3 candidates identified in both works.

•New discoveries:

- 4,975 new binary candidates (70.1%)
- 1,706 new triple-star candidates (89.6%)

•Total identified:

- 7,096 binary candidates
- 1,903 triple-star candidates

Observation Statistics

- Exposure distribution:
  - Single exposure: 108 binaries, 7 triple systems
  - Maximum exposures: 121 for binaries, 116 for triples
  - 6 exposures: 3,650 binaries, 1,312 triples
- Radial velocity characteristics:
  - Generally >60 km/s
  - Exponential distribution above 80 km/s

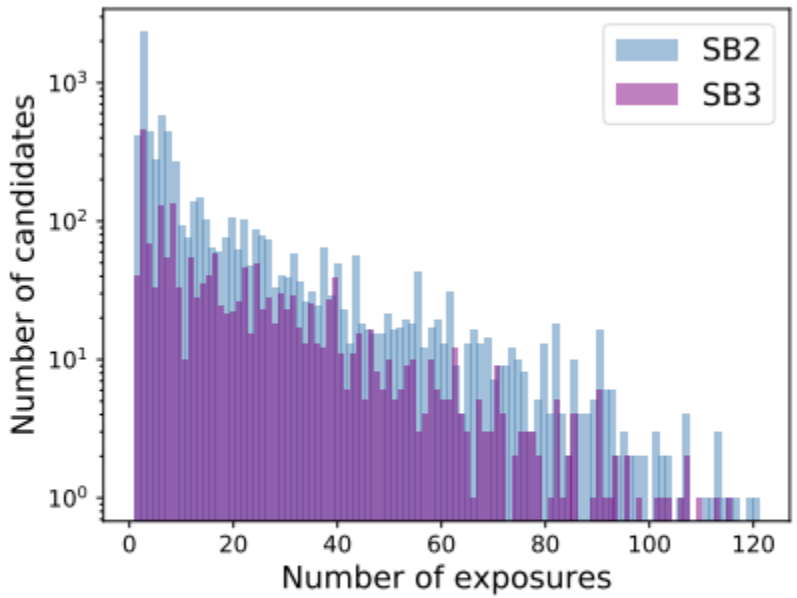

Figure 8. Number of detected SB candidates versus number of exposures (Blue arm). The number of candidates is in a logarithmic scale.

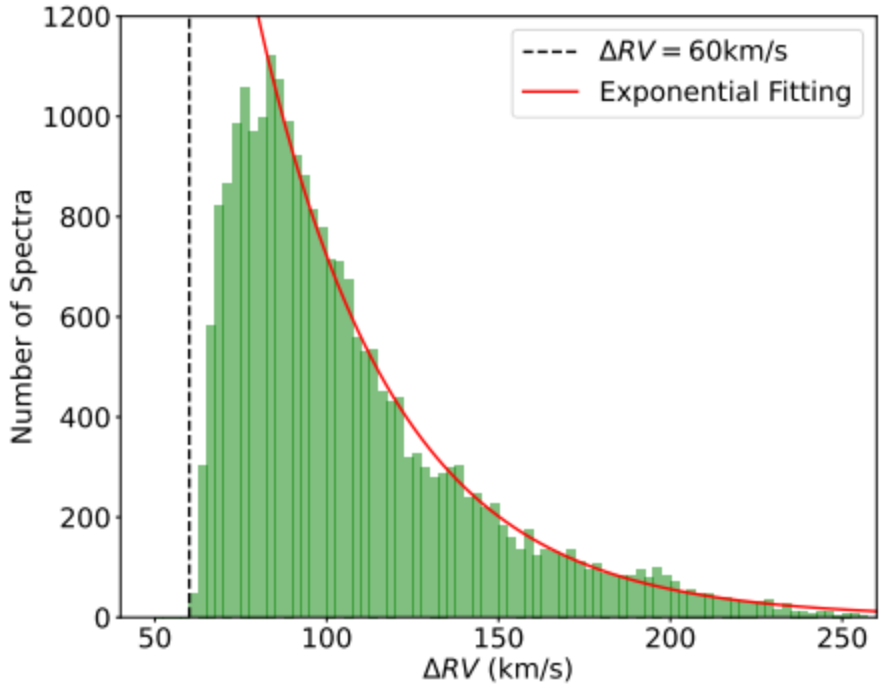

Figure 9. Distribution of RV differences (ΔRV) of all observed SB2 candidates in LAMOST-MRS DR9. The vertical dashed black line indicates the detection limit of ΔRV for SB2 in LAMOST-MRS spectra, which is about 60 km/s. The red line represents the exponential fitting curve. We applied an exponential function to fit the distribution of $\Delta RV > 80$ km/s.

DISCUSSION

Accuracy Enhancement

- Binary star detection:
  - Initial accuracy: 23.0%
  - Final accuracy: 99.7%
- Triple star detection:
  - Initial accuracy: 7.2%
  - Final accuracy: 26.3%
- Overall detection rate: 1.0% of selected dataset

Machine Learning Results

- Ensemble Learning Accuracy:
  - L2 (Binary): 99.7%
  - L3 (Triple): 19.4%
- Individual Classifier Performance:
  - Binary (C1/C2/C3): 71.6%/70.1%/44.8%
  - Triple (C1/C2/C4): 11.3%/10.5%/11.5%

Future Enhancement

- Improve triple-star system detection accuracy
- Optimize training data distribution
- Better velocity difference selection strategy
- Enhanced physical constraints
- Reduce false positives while maintaining high precision
- Address training sample impact on classification results

CONCLUSION AND PROSPECTS

Methodology: A hybrid human-AI approach comprising three key steps:

1.Traditional Cross-Correlation Function (CCF) analysis for initial component detection

2.Machine Learning methods with four DNN classifiers and ensemble learning

3.Human visual inspection for final verification

Major Findings:

•From LAMOST-MRS DR9 data (S/N > 5), we identified:

- 27,164 double-line spectra → 7,096 SB2 candidates
- 3,124 triple-line spectra → 1,903 SB3 candidates

•Significant new discoveries:

- 70.1% of SB2 candidates are newly identified
- 89.6% of SB3 candidates are newly identified

Method Improvements:

•Detection precision for SB2 candidates improved from 23.0% to 99.7%

•Combined CCF-ML approach significantly reduced time required for visual inspection

•Automated process effectively streamlines binary star detection in spectroscopic data

Future Work:

•Optimize classification efficiency and precision for SB3 candidates

•Enhance methodology through:

- Implementation of penalty parameters
- Application of multi-channel training
- Expansion of training sample size

•Analyze misclassification patterns from both CCF technique and ensemble learning method

Thanks!