Publications of the Astronomical Society of the Pacific, 135:084501 (15pp), 2023 August https://doi.org/10.1088/1538-3873/acea43

Identification of Blue Horizontal Branch Stars with Multimodal Fusion

Jiaqi Wei[1], Bin Jiang[1,*], and Yanxia Zhang[2,*]
[1] School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, 264209, Shandong, People's Republic of China; jiangbin@sdu.edu.cn
[2] CAS Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, 100101, People's Republic of China zyx@bao.ac.cn
Received 2023 February 24; accepted 2023 July 25; published 2023 August 2

Abstract

Blue Horizontal Branch stars (BHBs) are ideal tracers to probe the global structure of the milky Way (MW), and the increased size of the BHB star sample could be helpful to accurately calculate the MW's enclosed mass and kinematics. Large survey telescopes have produced an increasing number of astronomical images and spectra. However, traditional methods of identifying BHBs are limited in dealing with the large scale of astronomical data. A fast and efficient way of identifying BHBs can provide a more significant sample for further analysis and research. Therefore, in order to fully use the various data observed and further improve the identification accuracy of BHBs, we have innovatively proposed and implemented a Bi-level attention mechanism-based Transformer multimodal fusion model, called Bi-level Attention in the Transformer with Multimodality (BATMM). The model consists of a spectrum encoder, an image encoder, and a Transformer multimodal fusion module. The Transformer enables the effective fusion of data from two modalities, namely image and spectrum, by using the proposed Bi-level attention mechanism, including cross-attention and self-attention. As a result, the information from the different modalities complements each other, thus improving the accuracy of the identification of BHBs. The experimental results show that the $F1$ score of the proposed BATMM is 94.78%, which is 21.77% and 2.76% higher than the image and spectral unimodality, respectively. It is therefore demonstrated that higher identification accuracy of BHBs can be achieved by means of using data from multiple modalities and employing an efficient data fusion strategy.

Unified Astronomy Thesaurus concepts: Astronomy data analysis (1858); Classification (1907); Neural networks (1933)

1. Introduction

Blue horizontal branch stars (BHBs) are metal-poor (Santucci et al. 2015) A or B-type stars that burn helium in their cores (Barbosa et al. 2022). The inadequate understanding of the Galactic haloes' aggregate masses, dimensions, and formation history is primarily due to the inadequacy of extensive dynamic tracer sets at sufficiently substantial radii (Clewley et al. 2005). BHBs exhibit a luminosity that is both high and relatively constant (Barbosa et al. 2022). Specifically, their luminosity surpasses that of the majority of giant branch or Population II main sequence stars, and they exhibit distinctive spectral features that enable their identification (Smith et al. 2010). Due to their predictable brightness, they are often employed as standard candles to explore distant Galactic structures. Their use as tracers is widely sought in the studies concerning the kinematics and structural composition of our Galaxy (Vickers et al. 2021, 2012). Due to the considerable age of BHBs (Dotter et al. 2010), BHBs have become ideal for studying the structure of the older parts of the Galaxy (Culpan et al. 2021). Many works focus on using a growing sample of BHB halo tracers to probe the Milky Way's enclosed mass: Xue et al. (2008) used BHBs to derive precise constraints on the masses of Galactic dark matter halos; Gnedin et al. (2010) utilized 910 hypervelocity halo BHBs and blue straggler stars to map 80 kpc of the mass profile; Utkin & Dambis (2020) employed BHBs to simultaneously determine the distance from the Sun to the center of motion of the halo velocity field and a distance scale correction factor. In addition, BHBs have been applied to study dynamical substructures and stellar flows, anisotropy of the halo velocity distribution, etc. (Barbosa et al. 2022).

However, the main problem of BHBs as tracers are their relative sparsity compared to other tracers (e.g., turnoff stars) (Smith et al. 2010). Therefore, it is necessary to identify an increasing number of pure BHBs and thus obtain a more extensive sample of BHBs, which helps make further, more comprehensive studies of the Galaxy. The discrimination of BHBs from their main-sequence counterparts poses a critical

* Corresponding authors.

challenge in pursuing BHB identification (Culpan et al. 2021). The presence of a polluting population of high surface-gravity A-type stars and the blue stragglers makes identifying BHBs in distant halos unfeasible (Clewley et al. 2005). Traditional methods of identifying BHBs using spectroscopic data rely heavily on fitting Balmer line profiles that distinguish stars based on their surface gravity (Vickers et al. 2012, 2021). In addition to spectroscopic data, photometric methods can separate BHBs from other blue contaminants (Vickers et al. 2012). However, these traditional methods of identifying BHBs have the disadvantage of complex identification steps requiring more manual involvement. Furthermore, manual inspection has a subjective element, and the results may be accurate only to some extent.

As a result, the use of machine learning for the analysis and identification of celestial objects has come a long way facing the limitations of manual classification with respect to both efficiency and accuracy when massive amounts of observational data are obtained. In recent years, machine learning algorithms have also been applied to the identification of BHBs. Smith et al. (2010) studied the performance of some standard machine learning techniques (*k*-nearest neighbors, kernel density estimation and support vector machines) in identifying BHBs from photometric data. Vickers et al. (2021) used the XGBoost algorithm to identify BHBs from candidates, resulting in a BHB star catalog with a purity of about 86%. Identifying BHBs using machine learning methods does not require manual feature selection of the spectral data and can be done directly based on information from the entire spectrum. This can increase the spectral information used for classification while reducing manual involvement and improving classification efficiency, leading to more accurate classification results. In addition to classifying objects with spectral data, we can also classify objects using photometric images of the objects. Sky survey telescopes like the Sloan Digital Sky Survey (SDSS; York et al. 2000) already provide vast amounts of photometric data. It would be impractical for both individual researchers and the teams involved to examine all these images manually (Dieleman et al. 2015). With the development of deep learning (Lecun et al. 2015) techniques, convolutional neural networks (CNNs) are widely used in computer vision. Nowadays, more and more scientists also make use of CNNs for celestial object detection. For example, Aniyan & Thorat (2017) used CNNs to classify radio images of extended sources on a morphological basis; Pasquet et al. (2019) applied CNNs for photometric redshifts from SDSS images; Davies et al. (2019) utilized CNNs to identify gravitational lensing in astronomical images.

Multimodal learning is a general method for building artificial intelligence (AI) models that extract and correlate information from multimodal data (Baltrusaitis et al. 2019). Multimodal learning has been used in several areas (Khattar & Quadri 2022), such as visual question answering, emotion recognition, machine translation, cross-modal retrieval, and speech recognition. With the development of large survey telescopes, a massive amount of multi-source heterogeneous astronomical data, such as spectral and photometric data of astronomical objects, have been generated. These data can help us realize the classification, identification, and other studies of astronomical objects. However, existing techniques for classifying and identifying celestial objects based on deep learning often only use spectral or image data of celestial objects independently, and techniques for classifying objects with data from different modalities simultaneously through multimodal learning are still being explored. How the different modalities of astronomical data can complement each other to improve the efficiency and accuracy of classifying and identifying astronomical objects is the question that needs to be explored in this paper.

In multimodal learning, multimodal fusion (Baltrusaitis et al. 2019) is a widely studied topic, and it is very important to know how the information from different modalities can be adequately fused. Atrey et al. (2010) explored that the fusion of multiple modalities can provide complementary information and improve the accuracy of the overall decision process. So far researchers have proposed many methods to address this kind of problem. In this paper, we propose and implement a Bi-level attention mechanism-based Transformer multimodal fusion model called BATMM, with which automatic and efficient identification of BHBs can be achieved. The general workflow of the method is shown in Figure 1. First, the spectral and image data of objects are extracted by their respective encoders to obtain the features of the individual modes, followed by supervised training on the BHBs identification task using the Transformer fusion module we build. We improve the original self-attention mechanism in the Transformer by using the Bi-level attention mechanism. The improved model changes how the two modalities interact, which allows the data from the different modalities to be fused more effectively, thus improving the accuracy in detecting and identifying BHBs. We verify the model's performance by testing it on a test set.

This paper is organized as follows: Section 2 introduces our data sources. Section 3 illustrates the background of the current Transformer-based multimodal fusion approach. Section 4 describes the specific method used in this paper to fuse spectra and images. Section 5 shows our process for processing spectral and image data. Sections 6 and 7 present the evaluation metrics used for the experiments, the experimental results, the ablation experiments, and the comparison experiments. Section 8 discusses the experimental results, further model improvements, etc. Finally, Section 9 provides a summary of the paper.

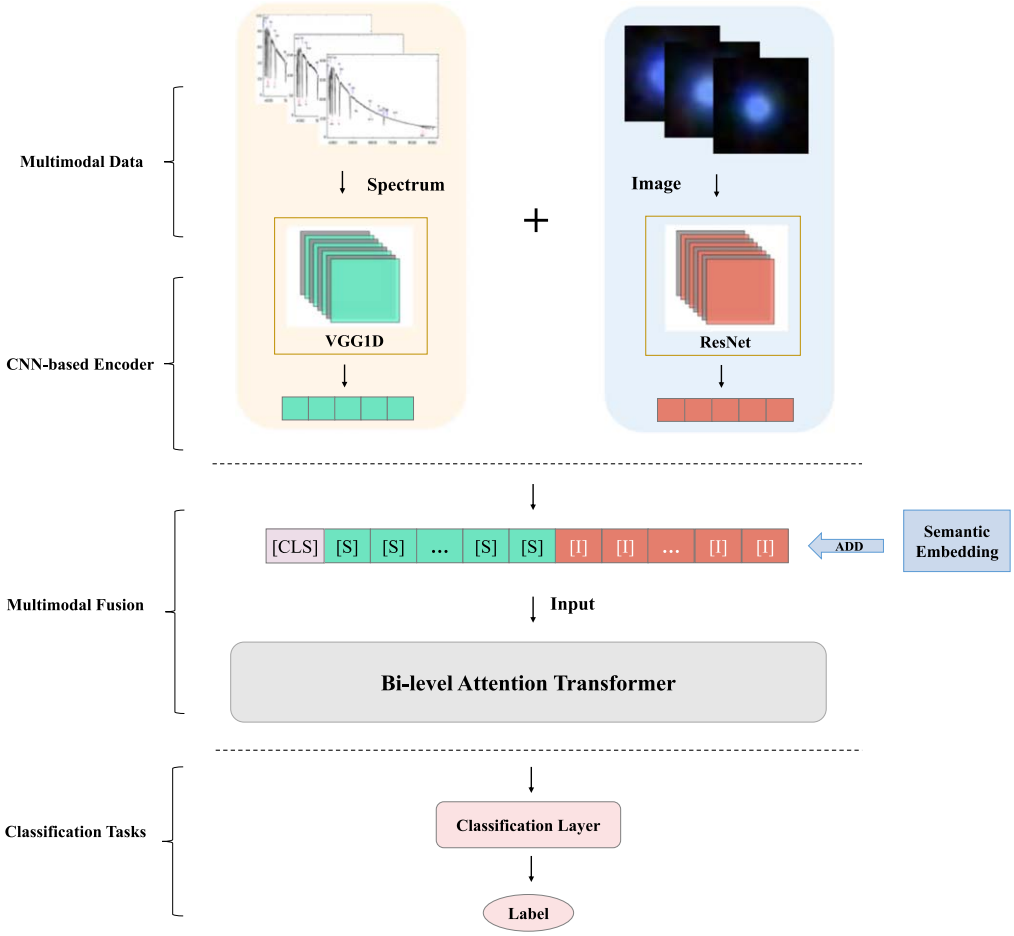

Figure 1. General workflow of the BATMM model.

2. Data

The SDSS (York et al. 2000) is a very successful large sky survey project, which performs imaging and spectroscopic survey. Data Release 16 (DR16; Ahumada et al. 2020) includes observations until 2018 August. It is the fourth data release of SDSS and contains different kinds of data, e.g., images, spectra and catalog data. DR16 keeps the version of DR13 (Albareti et al. 2017) imaging.

Based on spectra from SDSS, Xue et al. (2008) reprocessed a catalog of 10,224 BHB candidates through a dedicated hot star pipeline, identified 2558 BHBs and found less than 10% contamination, where the contaminants were mainly from main-sequence A-type (MSA) stars, in particular, the blue stragglers (BS). Xue et al. (2011) obtained a high probability sample of 4985 BHBs from SDSS DR8.

The selection of samples used in this study is presented in Table 1. The positive BHB sample is selected from Xue et al. (2011), which is considered representative (Bird et al. 2021; Vickers et al. 2021). In addition, we utilized negative samples of BS and A-type stars, sourced from Xue et al. (2008), due to their prevalent occurrence as primary contaminants employed for BHBs identification. We labeled the selected negative samples as non-BHBs. After collation, we select 4985 true BHBs and 7378 non-BHBs.

Table 1
Our Sample

Label	Type	Catalog
BHBs	BHBs	Xue et al. (2011)
Non-BHBs	BS/A-type stars	Xue et al. (2008)

Just for show, some spectra and their corresponding images of BHBs and non-BHBs are described in Figures 2 and 3, respectively.

3. Multimodal Fusion with the Transformer

Multimodal fusion is an important research direction in multimodal machine learning. In technical terms, multimodal fusion is the concept of integrating information from multiple

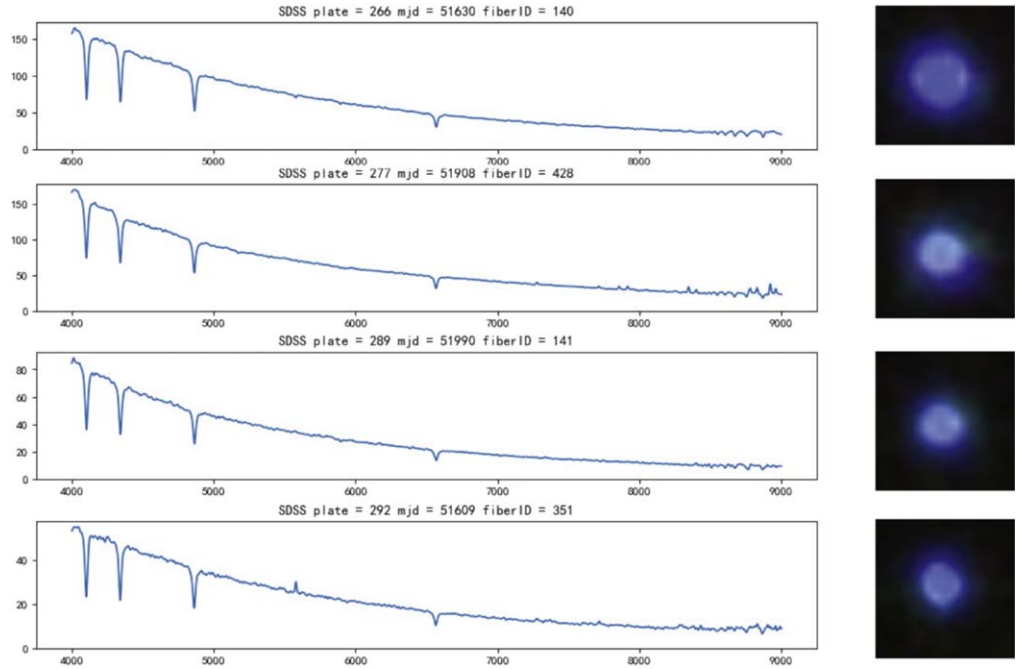

Figure 2. Some spectra of BHBs and corresponding images. The x-axis is is the observed wavelength in units of Å and the y-axis shows the flux density (f_λ) in units of 10^{-17} erg s^{-1} cm^{-2} Å^{-1}.

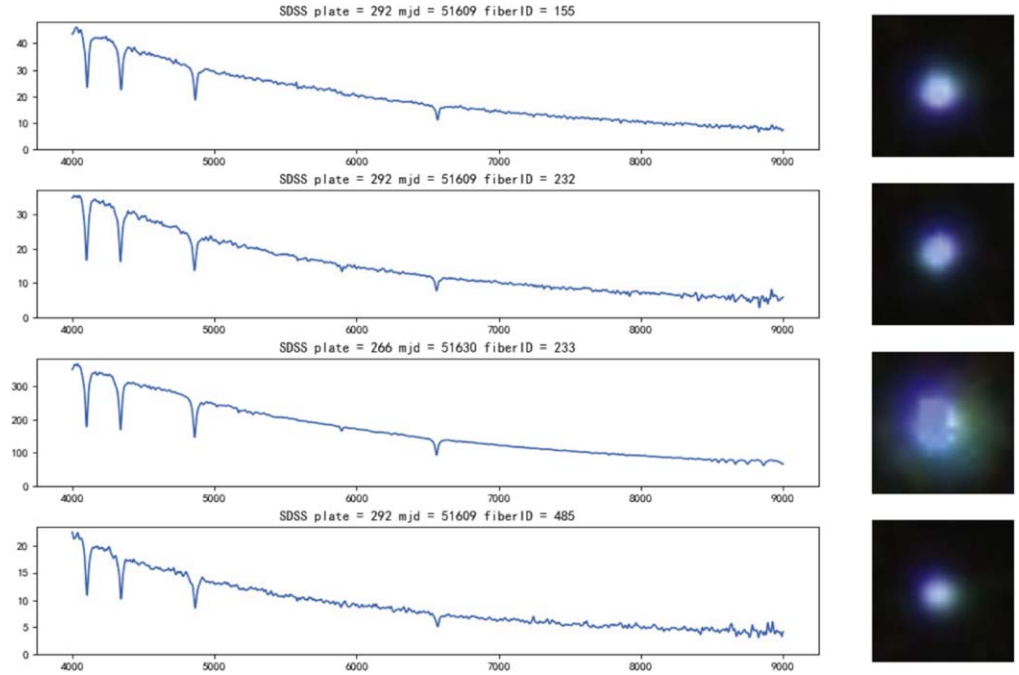

Figure 3. Some spectra of negative samples and their corresponding images. The x-axis is the observed wavelength in units of Å and the y-axis shows the flux density (f_λ) in units of 10^{-17} erg s^{-1} cm^{-2} Å^{-1}.

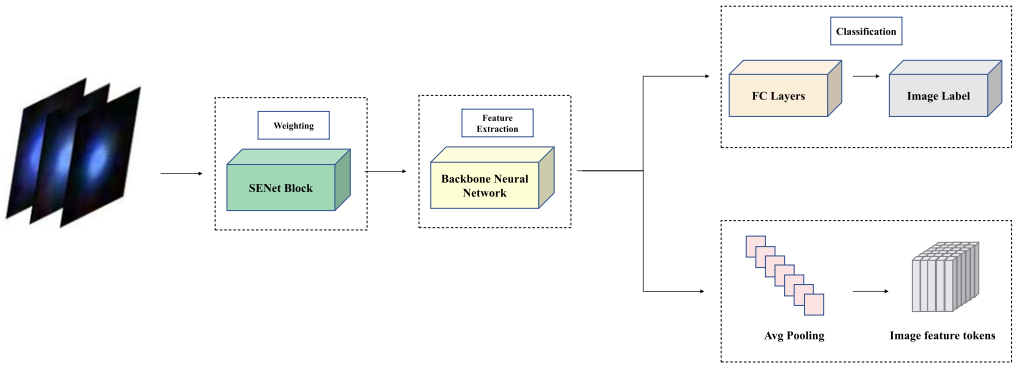

Figure 4. The workflow of image encoder.

modalities, often with the goal of predictive classification or regression (Baltrusaitis et al. 2019). There is a limit to how much information a single modality can represent, so researchers have been exploring ways to fuse information from multiple modalities.

The Transformer (Vaswani et al. 2017) differs from complex recurrent and convolutional neural networks in that the model structure of the Transformer uses an encoder–decoder framework, stacking multiple encoders and decoders to form the entire network, which completely avoids the use of convolution and recursion, and the entire network structure is composed entirely of self-attention mechanisms. Self-attention is an attention mechanism that associates different sequence positions to compute a uniform representation of the whole sequence. Compared with RNN (Borkowski et al. 2022), self-attention can improve parallelism and capture long-term dependencies better; compared with CNN (Lecun et al. 1998), which can extract local features, self-attention can model long-sequence remote relationships.

Considering the excellent results of Transformer, researchers have started investigating the use of Transformers for multimodal learning. The most significant advantage of the Transformer used for multimodal learning is its inherent strength and scalability in modeling various modalities and tasks (Xu et al. 2022). In the multimodal Transformer, the interaction between the different modalities is actually achieved through its internal attention mechanism. Therefore, Transformer-based multimodal learning can meet our needs for fusing multiple modalities.

4. Method

Spectral and photometric data are multimodal data: spectral data of celestial objects may provide celestial parameter measurement and are widely used by scientists to classify and study celestial objects; photometric images of celestial objects are easier to obtain than spectra and have very distinct visual characteristics, which are more intuitive and vividly helpful for the identification and classification of celestial objects. How to complement the data of these two modalities with each other and use the multifaceted and rich multimodal fusion data to classify BHBs is the issue to be solved in this paper. Transfer learning from representations of pre-trained models has been studied in many fields (Kiela et al. 2019). In this paper, we use an effective transfer learning strategy to extract multimodal features of celestial objects from spectral and image feature encoders as tokens input to a multimodal Transformer and then use our constructed multimodal Transformer fusion model for the BHB star sample to identify BHBs.

4.1. Image Encoder

Our astronomical images are from SDSS in five bands: u, g, r, i, z. We compare the five bands of data and align the five bands to form a five-channel matrix, which is our image data. We use residual neural network (ResNet; He et al. 2015) to extract visual features from celestial image data. Upon input of the image data into the encoder, a SENet Block (Hu et al. 2017) is incorporated to capture the interdependence among the five band photometric images. This is achieved by assigning weights automatically to each band, which facilitates the recognition of relevant image features by the model. At first, C calculates the channel features for global average pooling through Equation (1) where H and W represent the height and width of the image respectively, then the features are passed through two Fully Connected layers to model the correlation between channels through Equation (2) and output I_w as the weight of each band

$$C = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathrm{Img}(i, j) \tag{1}$$

$$I_w = \sigma(g(C, W)). \tag{2}$$

In this case, we extract the image features using only the backbone part of ResNet-50 and then flatten the features along the spatial dimension. The model structure diagram is shown in Figure 4. The flattened features are defined as $I = \{I_1, I_2, ..., I_k\} \in R^d$, where k is the number of pixels features, and d is the feature dimension of the pixel, which is set as 512 dimensions.

4.2. Spectrum Encoder

With the development of deep learning, CNNs are widely used to extract meaningful features from data, and the models applied for spectra have been shown to be effective for classification (Kim & Brunner 2017). Compared to traditional machine learning algorithms, deep learning techniques can automatically learn the underlying features from large and diverse data, eliminating the need for manual feature selection. In addition, CNNs can utilize the whole spectral information, thus avoiding the problems associated with human error when selecting profile ranges and shapes, and allowing more spectral data to be involved in classification and recognition (Vickers et al. 2021).

VGG (Simonyan & Zisserman 2014) is a very classical convolutional neural network classification architecture. The VGG1D (VGG based on one-dimension) convolutional model proposed in this paper is similar to the network structure of VGG-19, with the significant difference that the convolutional kernel we use is one-dimension so that the model can be applied to extract the features of the spectrum. Furthermore, we introduce a Self-attention Block to automatically assign weights to each wavelength range of the spectrum. This enhances the model's ability to emphasize the more informative wavelength ranges for the classification of BHBs during the training phase. Specifically, we need to cut the wavelength range of the spectra into the same window of 4000–9000 Å, thus ensuring that the spectral dimension of each object is 3522 dimensions. Following this step, the spectral data need to be divided into uniform sections. Upon manual calibration, it has been ascertained that the ideal sequence length for the encoding procedure is seven. As illustrated in Equation (3), the 3522 dimensional spectral data is divided into seven equal segments by manual adjustment, with the last data point rounded off. The segments are then concatenated vertically to form a spectral sequence of length 7, in which each segment possesses a dimensionality of 503, thereby achieving an equilibrium between expressiveness and manageability. The Q and K matrices are computed from the spectral sequence, as demonstrated in Equation (4), and the corresponding weights S_w are subsequently derived. This intervention has resulted in an enhanced effectiveness of the model during the training phase. The network architecture diagram is displayed in Figure 5

$$S = \{\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4, \hat{S}_5, \hat{S}_6, \hat{S}_7\} \tag{3}$$

$$S_w = \mathrm{softmax}\left(\frac{W_Q S \times (W_K S)^T}{\sqrt{d_k}}\right). \tag{4}$$

We can adjust the pooling and fully connected layers of the model after VGG1D to eventually map the extracted spectral features to 512 dimensions. After the adjustment, we are able to automatically extract the spectral features after feeding the spectral data into the model, which can then be fed into the subsequent Transformer network. The extracted spectral features are defined as $S = \{S_1, S_2, \ldots, S_k\} \in R^d$, where d is also set as 512 dimensions.

4.3. Multimodal Fusion

We use a modified Transformer-based module for the fusion of both image and spectral modalities at the feature level and for the identification of BHBs. Compared with early fusion and late fusion, our fusion strategy can achieve feature-level interaction and enable a more adequate fusion of multi-modality. The Transformer is currently shown to be compatible with a wide range of modalities. The Transformer is, therefore, compatible with this paper's spectral and image modalities. After obtaining the features of spectral and image pixels, we combine the features of the two modalities to construct a multimodal feature input sequence. We define two semantic embedding vectors, $\mathrm{Sem_S}$ and $\mathrm{Sem_I}$, to distinguish the spectral and image modalities, which have values of 0 and 1, respectively. We also add a special token [*CLS*] for learning the joint classification features, respectively. Finally, the Transformer's multimodal feature input sequence is formulated as follows:

$$\hat{I}_i = I_i + \mathrm{Sem_I},\ \hat{S}_i = S_i + \mathrm{Sem_S},\ i \in [1, k] \tag{5}$$

$$\mathrm{Input} = \{[CLS], \hat{I}_1, \hat{I}_2, \ldots, \hat{I}_k, \hat{S}_1, \hat{S}_2, \ldots, \hat{S}_k\}. \tag{6}$$

The Transformer's internal self-attention associates different positions of an input sequence to compute a uniform representation of the whole sequence. However, in order to make the model more applicable to our multimodal learning task, we improve the attention mechanism within the Transformer. We add a cross-attention mechanism that enables the spectral information of a celestial body to be focused differently on different regions of the image, resulting in a representation of the image associated with the celestial body's spectrum. We use cross-attention because when identifying BHBs by unimodality, the classification using the spectral modality alone is much better than the classification using the imaging modality, with approximately nearly 20% higher classification accuracy, due to the low differentiation of color features in the images. As a result, the two modalities are more differentiated, showing "strong modality" and "weak modality" respectively. If only the self-attention mechanism is used, it will lead to the model over-focusing on the information of the strong modality of the spectrum, resulting in less interaction between the two modalities and thus not fully utilizing the feature information of the weak modality of the image. Therefore, we hope that by using the cross-attention mechanism, the features of the spectrum can be used to highlight the features of the image regions associated with it, so as to enhance image features, thereby making full use of the information of the two modalities and improving the

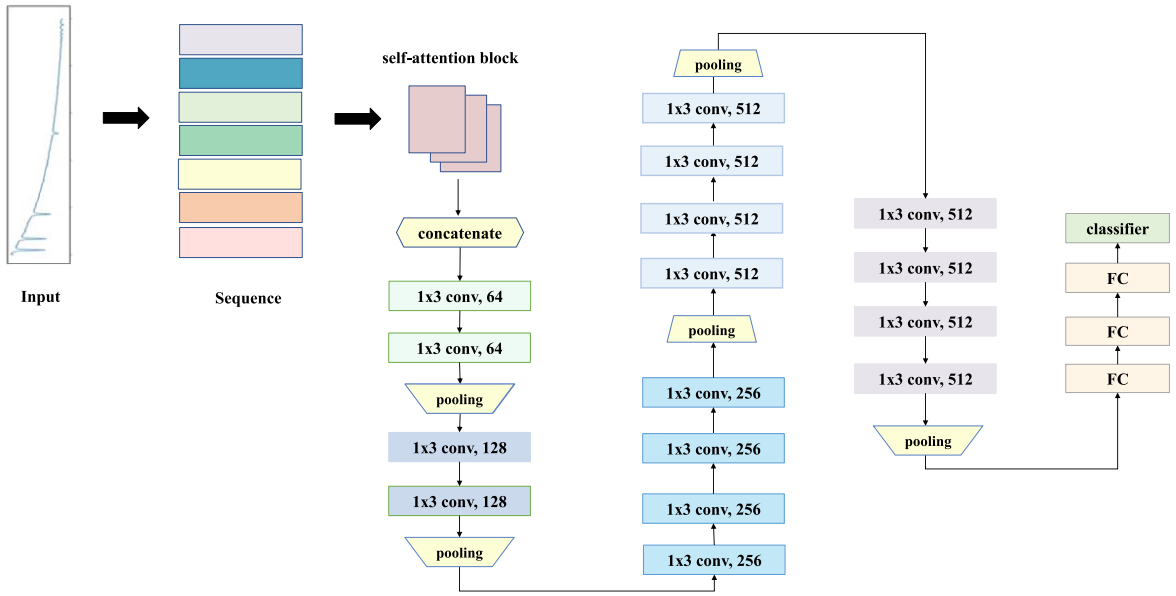

Figure 5. The workflow of VGG1D model.

performance of the multimodal classification model through more effective interaction and fusion of the two modalities.

As shown in Equation (7), we construct the query matrix ($W_{\mathrm{Q}}I$) from the feature vectors of the celestial image, the key matrix ($W_{\mathrm{K}}S$) and the value matrix ($W_{\mathrm{V}}S$) from the feature vectors of the spectrum, and then calculate the attention between the spectrum and different regions in the image by the query matrix and the key matrix. Among them, W_{Q}, W_{K} and W_{V} are three trainable weight parameters, S represents the extracted spectral features, I represents the extracted image features, and d_k represents the output dimension of $W_{\mathrm{K}}S$. Then, we have to use softmax normalization on the obtained attention A_{cross}, so that the sum of the weights is 1. Next, the feature representation of the image after cross-attention is obtained, as shown in Equation (8)

$$A_{\mathrm{cross}} = \mathrm{softmax}\left(\frac{W_{\mathrm{Q}}I \times (W_{\mathrm{K}}S)^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{7}$$

$$F_{\mathrm{h}} = A_{\mathrm{cross}} \times (W_{\mathrm{V}}S) \tag{8}$$

$$F_{\mathrm{cross}} = \mathrm{concat}(F_{\mathrm{l}},\cdots,F_{\mathrm{H}})\ W_{\mathrm{cross}}. \tag{9}$$

In practice, a multi-head cross-attention structure is often used, where we fuse the cascaded outputs from multiple attention sub-layers stacked in parallel through the projection matrix W_{cross} as shown in Equation (9), where each head F_{h} is calculated by Equation (8), $h \in [1, H]$, and W_{cross} is a linear projection matrix. By using the structure of the multi-head, it enables the model to process information from multiple representation subspaces simultaneously.

By performing cross-attention on the image modalities, we can obtain a relevant spectrum representation of the image features. As we can see from the way cross-attention is computed, cross-attention computes the image modal features under spectral conditions and is not able to perform cross-modal attention globally, thus losing the contextual information of both modalities. Therefore, after performing cross-attention on the image, we further connect the self-attention module to model the global feature representation of multiple modalities. In the self-attention mechanism, the query, key, and value vectors are generated for each token by multiplying the same input matrix with three trainable weight matrices W_{K}, W_{Q} and W_{V} for all multimodal feature tokens. We can implement self-attention according to Equations (10) and (11)

$$A_{\mathrm{self}} = \mathrm{softmax}\left(\frac{W_{\mathrm{Q}}M \times (W_{\mathrm{K}}M)^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{10}$$

$$F_{\mathrm{h}} = A_{\mathrm{self}} \times (W_{\mathrm{V}}M) \tag{11}$$

$$F_{\mathrm{self}} = \mathrm{concat}(F_{\mathrm{l}},\cdots,F_{\mathrm{H}})\ W_{\mathrm{self}} \tag{12}$$

$$F_{\mathrm{out}} = \mathrm{FFN}(F_{\mathrm{self}}) \tag{13}$$

where M represents all multimodal inputs after cross-attention, and d_k represents the output dimension of $W_{\mathrm{K}}M$. Then, we have to use softmax normalization on the obtained attention. Next, the computed attention weights A_{self} are multiplied by the value vectors $W_{\mathrm{V}}M$, and the feature representation F_{self} after self-attention is obtained, as shown in Equation (11). Based on Equation (12), we still use the multi-head attention mechanism to calculate the final result. The final output is calculated by a feed-forward neural network, where the FFN consists of a set of fully connected layers with ReLU activation functions.

The multimodal fusion module is shown in Figure 6. According to the previous section, we first obtain a multimodal

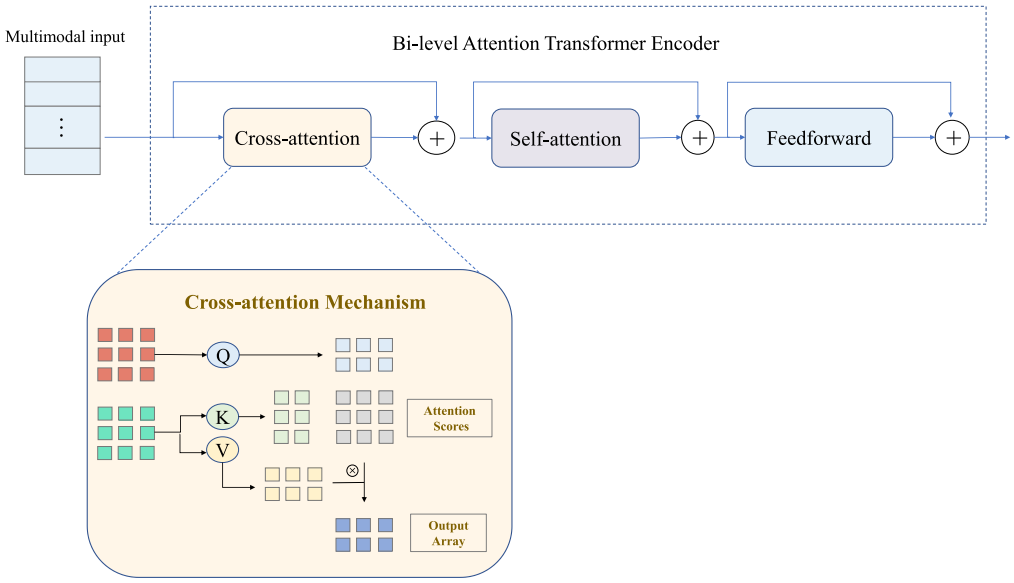

Figure 6. A Transformer encoder in the multimodal fusion module consists of cross-attention, self-attention, and feedforward layers.

representation of the data using the image encoder and the spectrum encoder and input the multimodal features into the Transformer at the same time. In the Transformer's attention module, by using cross-attention and self-attention mechanisms, data from different modalities can complement each other in terms of feature granularity, enabling feature-level fusion with potential semantics. Compared to the unimproved Transformer model, our feature-level fusion strategy allows for sufficient interaction between the more disparate modal features to form a unified representation of the fused modalities.

In summary, the BATMM multimodal model we have built consists of three components: an image encoder, a spectrum encoder, and a modified Transformer multimodal fusion module. First, we obtain the spectral and image feature representations through the spectrum and image encoders, then using a transfer learning strategy, the fusion module accepts the features from the spectra and images as input and uses our improved Transformer-based fusion model to perform deep information interaction and fusion of the features from the two modalities, thus improving the recognition performance of BHB stars. The entire model structure is shown in Figure 7.

5. Data Processing

According to the extinction map from SFD98 (Schlegel et al. 1998), we computed the extinction values of each BHBs in five bands (*ugriz*), employed statistical techniques to analyze the extinction values for the u, g, r, i, and z bands, and found that a significant proportion of objects in each band have extinction values that are in proximity to 0.1. Thus, we may disregard the correction values due to their trivial impact.

In order to ensure the validity of the data used for training and ensure the proper training of the subsequent network and accelerate the convergence of the model, we first need to pre-process the acquired image data and the spectral data separately. The number of samples in the unprocessed data set, the number of pre-processed image samples and the number of pre-processed spectral samples are listed in Table 2.

5.1. Image Processing

First, we enhance the quality of our BHBs image data set by removing images with quality problems due to inaccurate band alignment, etc., so that subsequent models can learn more accurate and representative image features. After removing, we need to normalize and standardize the remaining images in order to speed up the training of the model. We can use Equation (14) to normalize the pixel data of the image and Equation (15) to standardize the data

$$I_i = \frac{I_i - I_{\min}}{I_{\max} - I_{\min}} \tag{14}$$

$$I_i = \frac{I_i - \mu}{\sigma} \tag{15}$$

where I_i represents the pixel point to be processed, $I_{\max}$ is the maximum value among all the pixel data, $I_{\min}$ is the minimum value among all the pixel data, μ is our defined pixel mean and σ is our defined pixel standard deviation, equal to the mean and standard deviation values of the ImageNet database, respectively. After normalization, the image pixel values are adjusted to the interval [0, 1]. After standardization, the image pixel values are transformed into a distribution of mean and standard deviation values of the ImageNet database.

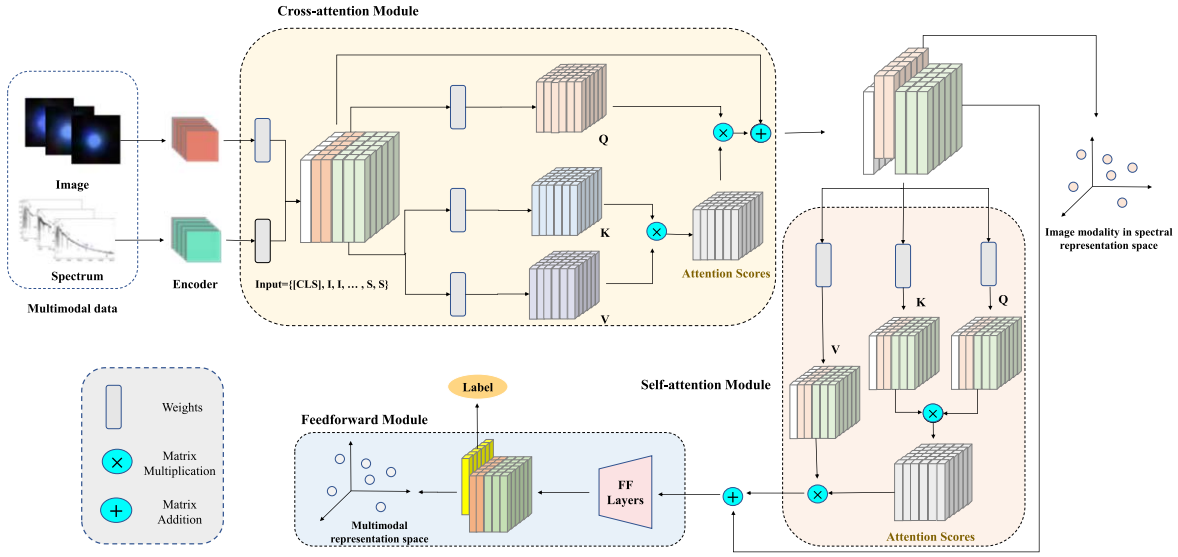

Figure 7. Throughout our multimodal fusion process, this diagram depicts the entire detailed process of multimodal data from spectrum and image from the data input, to the final output. First, the pre-processed multimodal data are fed separately into the encoder model for multimodal feature extraction. In this case, spectral features are extracted from the spectral data by our defined VGG1D convolutional encoder, and image features are extracted from the image by the ResNet convolutional model. We then feed the extracted features into our improved Transformer multimodal fusion network simultaneously, and by using the cross-attention and self-attention modules, we can achieve a feature-level fusion with potential semantics between the two modalities, allowing sufficient interaction between the more divergent modal features to form a fused one. Finally, a unified representation of the two modalities is formed after the feedforward layers. Finally, the fused features are classified using a classifier to obtain the prediction results of the data.

Table 2
The Number of Samples in our Data Set Before and After Data Processing

Type	Original	After Image Processing	After Spectrum Processing
BHBs	4985	4810	4752
Non-BHBs	7378	7378	7280
Total	12,363	12,188	12,032

5.2. Spectrum Processing

Our spectral data also come from SDSS DR16. First, to be able to feed the spectral data into our deep learning model for subsequent training, we need to cut the wavelength range of the spectra into the same window of 4000–9000 Å, thus ensuring that the spectral dimension of each object is 3522 dimensions. In this process, some spectra have less than 3522 dimensions in the wavelength range we divide, and cannot be used as part of the data set for subsequent training, so we remove this part of the spectra. At the same time, to ensure the one-to-one correspondence between the spectra and the two modalities of the images, we remove the images corresponding to the spectra that need to be removed as well. After the deletion, our multimodal data set has 4752 BHBs samples and 7280 non-BHB samples, and the ratio of positive and negative samples is about 65%.

Then, we need to pre-process the spectra for normalization (Paoletti et al. 2018). Since the fluxes of the same spectral data on different bands tend to have different scales, we first scale the data to the same distribution by normalization measures. We need to normalize the scaling of all valid spectral data using the following Equation (16)

$$S_i = \frac{S_i - S_{\min}}{S_{\max} - S_{\min}} \tag{16}$$

where S_i is the spectral data to be processed, $S_{\max}$ is the maximum value in that spectral data and $S_{\min}$ is the minimum value in that spectral data. After this normalization process, all the spectral data are scaled to the distribution of (0, 1], thus ensuring that they can be trained effectively using the model we built.

6. Experiment

We randomly take 70% of the samples for training, 20% for validation, and 10% for testing. That is, to ensure the reliability of our model's performance, we divide the data set into a training set and an independent test set, with the former consisting of 90% of the data and the latter comprising the remaining 10%. Specifically, the training set was utilized for the training process, while the test set was kept separate for the sole purpose of evaluating the model's performance in an unbiased manner.

Table 3
Experimental Environment

Placements	Configuration
GPU	NVIDIA GeForce RTX 3070
Computer language	Python
Python editor	PyCharm
Function library	Pytorch 1.10.0 CUDA 11.1

After extracting the features of spectra and images using the encoder, the modified Transformer-based multimodal fusion module described in Section 4.3 is used to combine the features of both modalities. We define 8 layers of transformer blocks, each with 8 attention heads, which are empirical values. For large networks such as Transformer, the model is unstable in the initial stage of training and can easily become difficult to train. Specifically, we employ AdamW with an initial learning rate of 1e-4 and weight decay of 5e-2 as the optimizer for the Transformer. We still use the Cross-Entropy Loss Function and a dynamically adjusted learning rate strategy to update the model's gradient. We first change the model's hyper-parameters on the validation set, then complete the model's training. Finally, the model is loaded with the saved optimal model parameters and weights and tested on the defined test set.

Our hardware and software experimental environments are shown in Table 3 below. In order to speed up the processing of image and spectral data, we use NVIDIA GeForce RTX 3070 graphics card to speed up the data, thus significantly reducing the time required to train our model.

7. Model Evaluation and Experimental Results

We first define the evaluation metrics for the model in Section 7.1. We then show how well our multimodal fusion model performs on the test set in Section 7.2. To demonstrate the effectiveness of our improvements to the multimodal fusion module, we perform ablation experiments in Section 7.3. Finally, we compare results with several powerful baselines in Section 7.4. To demonstrate the validity of multimodal learning, we compare the performance of the multimodal model with two unimodal models using only images and only spectra; to show the superior performance of our proposed multimodal fusion model, BATMM, we compare it with other common and effective baseline models. To ensure the validity and reliability of our experimental results and to reduce the instability of the model performance due to random factors during the training process, we run five times randomly and calculate the average results with the standard deviation values for each experimental task, thus greatly improve the reliability of our model.

7.1. Evaluation Metrics

In this experiment, we use a confusion matrix to describe the classification results based on the actual class of each object and the classes predicted according to our model. Table 4 shows the confusion matrix, where each row of the confusion matrix represents the real class, and each column represents the predicted class. The positive and negative examples in this paper represent the BHBs and non-BHBs, respectively. TP denotes the number of true BHBs predicted as BHBs, FN denotes the number of true BHBs incorrectly predicted as non-BHBs and FP denotes the number of non-BHBs incorrectly predicted as BHBs.

In this study, since BHBs recognition is a binary classification problem, we use Precision, Recall, and $F1$ score, which are commonly used in binary classification problems, to evaluate the performance of the proposed method. Precision is the proportion of all predicted positive samples that are true positive; Recall is the proportion of all positive samples that are correctly predicted as positive. Precision and Recall are defined by using Equations (17) and (18), respectively. $F1$ score is computed by Equation (19) and interpreted as the summed average of Precision and Recall. Precision, Recall and $F1$ score are metrics used to evaluate the performance of a model. The larger Precision, Recall and $F1$ score of a model, the better performance the model has

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{18}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{19}$$

For machine learning, the Receiver Operator Characteristic (ROC) curves are widely used in binary classification problems to evaluate the confidence of classifiers. In addition, P–R curves are also used to evaluate the classification performance of models. In the ROC curve, the x-axis is FPR and the y-axis is TPR. FPR refers to the probability that actual negative samples are incorrectly predicted as positive samples, defined as Equation (20). TPR refers to the probability that the actual positive samples are correctly predicted, defined as Equation (21). In the P–R curve, the x-axis represents Recall and the y-axis shows Precision. AUC is defined as the area under the ROC curve. When AUC is closer to 1, the model has better performance. The ROC curve is near the left upper angle and the P–R curve is near the right upper angle, the model has better classification accuracy

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{20}$$

Table 4
Confusion Matrix

	Predicted as Positive Examples	Predicted as Negative Examples
True positive examples	TP	FN
True negative examples	FP	TN

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \quad (21)$$

7.2. Experimental Results

The final saved parameters and weights are loaded into the model for evaluation. To evaluate the validity of our proposed model, a pre-preserved and independent test set is utilized. This test set is not used during the training phase. As shown in Table 5, for the whole sample, Precision, Recall and $F1$ of our model are 94.45%, 95.18% and 94.78%, respectively; for BHBs, these metrics are 91.58%, 95.78% and 93.63%, separately. The results all show that the BATMM model has satisfactory classification performance.

We plot the ROC curve and PR curve for one of the times, as shown in Figure 8. The *P–R* curves are plotted separately for BHBs, non-BHBs and the whole sample. In the ROC curve, AUC is 0.99, as well as the curve positions in the two curve diagrams, which further indicates that the BATMM model has an excellent performance.

By incorporating Slef-attention and SENet Block modules, the encoder is able to dynamically allocate weights to various features throughout the training phase. Figure 9 depicts the encoders' assigned weights. Upon examining the weights assigned to the seven spectral wavelength ranges, it was discovered that the blue range of spectra plays a more vital role in classifying BHBs. Similarly, the analysis of the image encoder weights for each band indicated that the *u*-band photometry was crucial for the classification of BHBs, contributing approximately 32%. Based on these results, we can infer that the *u*-band or blue range of the spectra are pivotal attributes that facilitate the differentiation of BHBs from other stars. It is noteworthy that BHBs display a sharp and prominent Balmer jump in this region. Therefore, a reasonable increase in the weighting of these inputs can enhance the performance of our model.

7.3. Ablation Study

To validate the effectiveness of the image and spectrum encoders, we conduct a series of ablation experiments to examine the impact of assigning weights to different wavelength ranges of the spectrum and image on the classification accuracy of the model. The experimental results are shown in Table 6. Our findings indicate that the incorporation of a Self-attention Block in the spectrum encoder allows for distinct weights to be assigned to diverse wavelength ranges, thereby enhancing the model's classification accuracy. Furthermore, the addition of the SENet Block to the images results in improved classification performance, demonstrating the effectiveness of the proposed encoder enhancements.

Table 5
Experimental Results of the BATMM Model

Type	Precision($\% \pm \sigma$)	Recall($\% \pm \sigma$)	$F1$ Score($\%\pm\sigma$)
BHBs	91.58 ± 0.95	95.78 ± 0.19	93.63 ± 0.52
Non-BHBs	97.33 ± 0.19	94.59 ± 0.48	95.94 ± 0.25
ALL	94.45 ± 0.45	95.18 ± 0.29	94.78 ± 0.38

The Transformer can process the spectra after encoding alone, then we verify whether the joining of image modalities is helpful. When using Transformer to fuse multimodal data, we assign one semantic embedding to the image and another to the spectrum in order to distinguish the features of the two modalities. Since the image is a weak modality compared to the spectrum, in order to prevent the image from negatively affecting the multimodal fusion and thus reducing the fusion performance, and allowing the model to learn more effective multimodal global features, we add the cross-attention module to the Transformer internally to first obtain the image representation related to the spectrum and then use self-attention to obtain the global representation after multimodal fusion. Therefore, to prove whether our improvements to Transformer are effective, we perform an ablation study on the incorporation of semantic embedding and cross-attention.

All experimental data sets and hyper-parameter configurations are the same as before, and the experimental results are shown in Table 7. Compared with using Transformer alone to process the spectra, the classification performance of the model has no significant variation after adding the images without the cross-attention module, and it may even be reduced. Based on both spectra and images, the performance of the model is improved when adding semantic embedding or the cross-attention module, and the classification performance of the model is further improved when adding semantic embedding and the cross-attention module, which demonstrates the validity of the model improvement using cross-attention. Thus, the contribution of cross-attention to the recognition of BHBs is verified by obtaining the image representation related to the spectra and then performing multimodal fusion, which enables

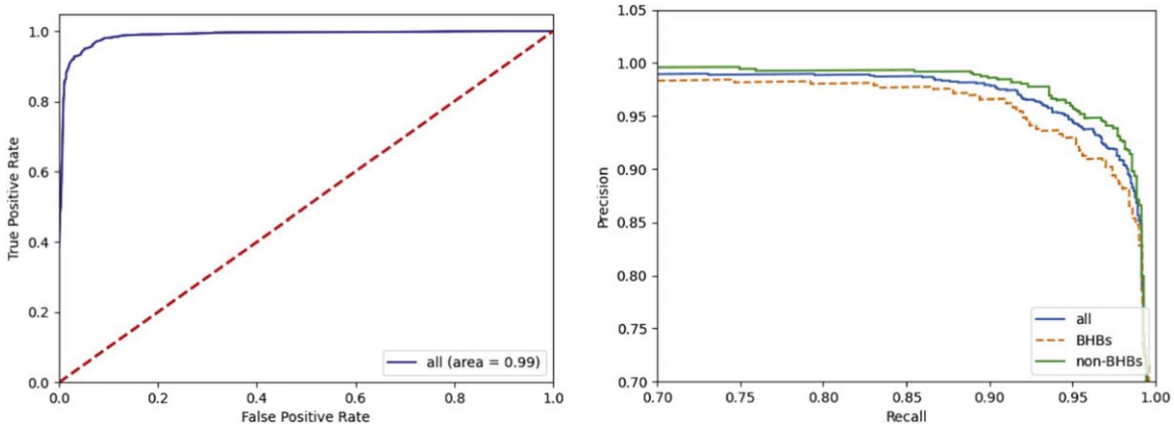

Figure 8. The ROC curve (left panel) and *P*–*R* curve (right panel).

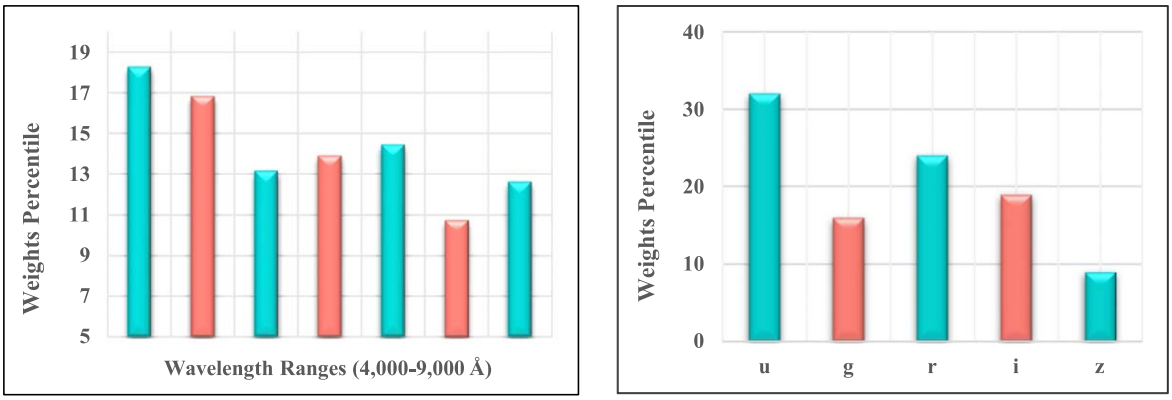

Figure 9. The weights assigned by the encoders.

Table 6
The Results of Ablation Experiments for Different Models

Modality	Model	Precision ($\% \pm \sigma$)	Recall ($\% \pm \sigma$)	*F*1 Score ($\% \pm \sigma$)
Spectrum	VGG1D	91.15 ± 1.08	91.05 ± 1.41	91.00 ± 1.12
	+ self-attention	92.18 ± 0.39	91.93 ± 0.51	92.02 ± 0.20
Image	ResNet	72.58 ± 1.09	72.60 ± 1.09	72.57 ± 1.09
	+ SENet Block	72.96 ± 1.25	73.12 ± 0.88	73.01 ± 1.10

more efficient interaction and fusion of the two modalities to improve the classification performance.

7.4. Model Comparison and Analysis of Results

For the BATMM multimodal fusion model, we first compare the classification performance with the unimodal model to prove the superior performance of our proposed multimodal learning strategy for the identification of BHBs. Second, we also compare it with other efficient fusion strategies. To ensure the validity of the comparison results, the simple linear classifier is used for each baseline during classification. We detail each comparison model below. The comparison results are shown in Table 8 and Figure 10.

7.4.1. Image-only

We use a pre-trained model of ResNet-50 and pool the model output at the end of the model to generate a 512 dimensional feature vector for each image, and then apply a linear classifier to classify the image data.

Table 7
The Results of Ablation Experiments of Transformer (Sem refers to Semantic Embedding, CA represents Cross-attention)

Model	Precision($\% \pm \sigma$)	Recall($\% \pm \sigma$)	$F1$ Score($\% \pm \sigma$)
Transformer(with spectrum)	93.84 ± 0.22	93.78 ± 0.11	93.80 ± 0.06
+ Image	93.66 ± 0.67	93.50 ± 0.69	93.58 ± 0.69
+ Image + Sem	93.73 ± 0.39	94.23 ± 0.12	93.95 ± 0.17
+ Image + CA	94.27 ± 0.83	94.45 ± 0.83	94.35 ± 0.83
+ Image + Sem + CA(BATMM)	94.45 ± 0.45	95.18 ± 0.29	94.78 ± 0.38

7.4.2. *Spectrum-only*

We utilize the VGG1D convolutional model proposed to acquire spectral features. Then a 512 dimensional representation vector is generated for each spectrum by pooling operations and mapping of fully connected layers, and a linear classifier is performed for classification. We compare our proposed model with other models used for the identification of BHBs. Previous studies in the field of BHBs identification using machine learning have primarily utilized support vector machines (SVMs; Smith et al. 2010) and XGBoost (Vickers et al. 2021) models, with XGBoost exhibiting a superior recognition accuracy in comparison.

7.4.3. *Early Add*

In practice, the early add is a simple and effective way of multimodal fusion, with the advantage of low complexity (Gavrilyuk et al. 2020). We sum the feature vectors generated for each spectrum and image with Spectrum-only and Image-only, resulting in a 512 dimensional feature vector, which we treat as a multimodal feature vector, and use a linear classifier for classification.

7.4.4. *Early Concat*

For multimodal methods, concatenation is often used as an essential baseline (Sun et al. 2019; Shi et al. 2022). We concatenate the feature vectors generated for each spectrum and image with Spectrum-only and Image-only using a concatenated feature dimension of 1024 dimensions and then use a linear classifier for classification.

7.4.5. *VGG1D*

CNNs are very powerful in feature extraction. We first concatenate the spectral and image feature vectors to produce a multimodal feature encoding with a dimension of 1024, then feed it into the VGG1D convolutional network we have built to extract multimodal features and realize the classification task with these extracted features.

As indicated in Table 8, comparing Early Add, Early Concat, VGG1D, and our proposed BATMM multimodal models with the Spectrum-only and Image-only unimodal models, it is seen that the $F1$ scores of our multimodal models after fusing spectral and image information are higher than those of the two unimodal models. In particular, our proposed BATMM multimodal classifier has $F1$ score of 94.78%, which is 2.76% and 21.77% higher than the two unimodal models respectively. It is also noticed that there is a difference between the classification performance using only images and only spectra, and the performance with only spectra is much better than that with only images, mainly because spectra have more abundant information (e.g., spectral line strength and width, different continuum shapes) than images and the main features of images (e.g., color, edges, and luminosity) vary less between positive and negative samples. In contrast, we obtain better classification results by fusing spectral and image information to complement each other. As a result, it is further proved that the accuracy of classification and recognition of BHB stars through multimodal learning is effectively improved.

Our proposed model was compared to previous models, namely SVM and XGBoost, that have utilized machine learning techniques for BHB identification. The results in Table 8 showed that our model outperformed both SVM and XGBoost models, both in the case of using only spectral data through the VGG1D model and multimodal data, implying that our proposed model has superior performance.

As shown in Figure 10, comparing the proposed BATMM model through the fusion strategies with Early Add, Early Concat, and VGG1D, the BATMM classifier shows the best performance in terms of Precision, Recall and $F1$ score, which indicates that the performance of multimodal classification models may be improved by implementing a potential feature-level interaction between two modalities through the Transformer and introducing an attention mechanism that enables fuller fusion between the features of the more divergent modalities.

8. Discussion

The experimental results show that the proposed BATMM multimodal fusion model has superior performance compared to unimodal classification models and other multimodal classification models. Therefore our model can be used to identify BHBs more accurately. With deep learning algorithms, we can automate the classification of celestial objects with higher accuracy when faced with large amounts of astronomical data, thus freeing us from the tedious steps of manual identification of BHBs.

Our multimodal model uses the transfer learning strategy where the multimodal fusion module and the feature extraction module do not have to be trained simultaneously, thus allowing flexibility to replace encoders for better results when more powerful encoders for images or spectra are available. In

Table 8
Performance Comparison of the Improved BATMM Model and the Other Models

Input	Model	Precision ($\% \pm \sigma$)	Recall ($\% \pm \sigma$)	$F1$ Score ($\% \pm \sigma$)
Image-only	ResNet	72.96 ± 1.25	73.12 ± 0.88	73.01 ± 1.10
Spectrum-only	SVM	75.80 ± 0.49	89.41 ± 0.32	82.04 ± 0.33
	XGBoost	84.46 ± 0.38	88.91 ± 0.53	86.63 ± 0.19
	VGG1D	92.18 ± 0.39	91.93 ± 0.51	92.02 ± 0.20
Multimodal	Early Add	92.59 ± 0.34	91.99 ± 0.24	92.25 ± 0.25
	Early Concat	93.12 ± 0.70	92.49 ± 0.76	92.77 ± 0.72
	VGG1D	94.15 ± 0.61	93.81 ± 0.59	93.96 ± 0.57
	BATMM	94.45 ± 0.45	95.18 ± 0.29	94.78 ± 0.38

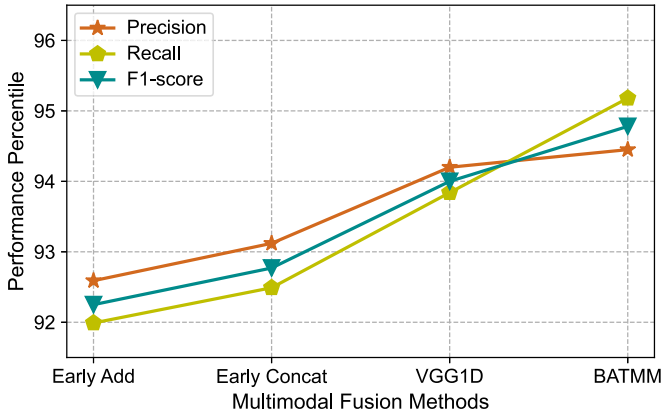

Figure 10. Comparison of multimodal fusion methods.

addition, our model is well scalable and can be easily extended to other applications for identifying and classifying target objects. Given the limitations of supervised learning, when our model needs to be applied to the classification task of other celestial objects, we need to reacquire the multimodal data set corresponding to the images and spectra of the objects and retrain the model using the new data set. When dealing with different data, we also need to consider the datas characteristics, so we can suggest other strategies to improve the model and make it more customizable.

For both modalities, we can do a further fine-grained alignment. For example, we can add a comparison learning strategy in a subsequent improvement. Before the modalities are fused, we first realize the alignment between the different modalities through contrast learning to further close the distance between the two modalities, so that the different modalities can interact more fully at a fine-grained level in the process of multimodal fusion. We can also explore other ways to improve the Transformer's attention mechanism to make it more applicable to a wide range of multimodal learning tasks.

In contemporary times, large survey telescopes generate copious amounts of celestial data. Simultaneously, the immense quantity of data requiring analysis and processing presents a considerable challenge. Consequently, the development of accurate and efficient deep learning models for the automated examination and study of celestial objects is a topic deserving of ongoing investigation. Enhancements in the precision and dependability of these models can be achieved by combining heterogeneous, multi-source, and multimodal data, thereby optimizing the utilization of the abundant celestial observational data available. As such, it is essential to consider the expansion of these models. In addition to the dual modalities of image and spectral data for celestial entities, efforts should be made to integrate information from additional modalities, as well as combine data from different wavelengths or telescopes.

9. Conclusion

We introduce the BATMM multimodal learning model designed for BHBs identification, comprising three modules: the spectrum encoder, the image encoder, and the Transformer multimodal fusion module. To improve the identification accuracy of BHBs using image and spectral features, we allocate distinct weights to each feature and find that the blue range of the spectrum and u-band of the image are crucial. Furthermore, we incorporate a cross-attention mechanism in the Transformer multimodal fusion module, enhancing the attention ability of the original model based on the characteristics of the two modalities. Our proposed model is thus more suitable for multimodal classification tasks.

Based on spectra and images, various models are compared, of which our proposed BATMM model shows its superiority and achieves 94.45% Precision, 95.18% Recall and 94.78% $F1$ score on the test set. The results show that our proposed multimodal learning strategy has excellent performance and significantly improves the classification accuracy compared to unimodal models. Moreover, the BATMM fusion model performs better than other multimodal fusion strategies. In addition, the effectiveness of our improvements is demonstrated through ablation experiments. Finally, the BATMM multimodal learning model can be used to automatically identify BHBs, which can help construct a larger sample of

BHBs and help study the Galaxy even further. This model may also be applied to other classification problems faced in astronomy utilizing multimodal data.

Acknowledgments

We thank the referee very much for their constructive comments and suggestions. The study was funded by the National Natural Science Foundation of China under grant Nos. 12273076, 12133001 and 11873066, the China Manned Space Project with science research grant Nos. CMS-CSST-2021-A04 and CMS-CSST-2021-A06, and the Natural Science Foundation of Hebei Province under grant No. A2018106014. We acknowledge SDSS databases. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS website is www.sdss.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

Data Availability

The data underlying this article are available in SDSS, at https://dr16.sdss.org/ (catalog data).

ORCID iDs

Jiaqi Wei https://orcid.org/0000-0002-8802-2241
Bin Jiang https://orcid.org/0000-0002-2897-5745
Yanxia Zhang https://orcid.org/0000-0002-6610-5265

References

Ahumada, R., Allende Prieto, C., Almeida, A., et al. 2020, ApJS, 249, 3
Albareti, F. D., Allende Prieto, C., Almeida, A., et al. 2017, ApJS, 233, 25
Aniyan, A. K., & Thorat, K. 2017, ApJS, 230, 20
Atrey, P. K., Hossain, M. A., El-Saddik, A., & Kankanhalli, M. S. 2010, Multim. Syst., 16, 345
Baltrusaitis, T., Ahuja, C., & Morency, L. 2019, ITPAM, 41, 423
Barbosa, F. O., Santucci, R. M., Rossi, S., et al. 2022, ApJ, 940, 30
Bird, S. A., Xue, X.-X., Liu, C., et al. 2021, ApJ, 919, 66
Borkowski, L., Sorini, C., & Chattopadhyay, A. 2022, CoStr, 258, 106678
Clewley, L., Warren, S. J., & Hewett, P. C. 2005, MNRAS, 362, 349
Culpan, R., Pelisoli, I., & Geier, S. 2021, A&A, 654, A107
Davies, A., Serjeant, S., & Bromley, J. M. 2019, MNRAS, 487, 5263
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Dotter, A., Sarajedini, A., Anderson, J., et al. 2010, ApJ, 708, 698
Gavrilyuk, K., Sanford, R., Javan, M., & Snoek, C. G. M. 2020, arXiv:2003.12737
Gnedin, O. Y., Brown, W. R., Geller, M. J., & Kenyon, S. J. 2010, ApJL, 720, L108
He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv:1512.03385
Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. 2017, arXiv:1709.01507
Khattar, A., & Quadri, S. M. K. 2022, IEEE Access, 10, 92889
Kiela, D., Bhooshan, S., Firooz, H., Perez, E., & Testuggine, D. 2019, arXiv:1909.02950
Kim, E. J., & Brunner, R. J. 2017, MNRAS, 464, 4463
Lecun, Y., Bengio, Y., & Hinton, G. 2015, Natur, 521, 436
Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proc. IEEE, 86, 2278
Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. 2018, JPRS, 145, 120
Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26
Santucci, R. M., Beers, T. C., Placco, V. M., et al. 2015, ApJL, 813, L16
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Shi, B., Hsu, W.-N., Lakhotia, K., & Mohamed, A. 2022, arXiv:2201.02184
Simonyan, K., & Zisserman, A. 2014, arXiv:1409.1556
Smith, K. W., Bailer-Jones, C. A. L., Klement, R. J., & Xue, X. X. 2010, A&A, 522, A88
Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. 2019, arXiv:1904.01766
Utkin, N. D., & Dambis, A. K. 2020, MNRAS, 499, 1058
Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, in NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, ed. U. von Luxburg et al. (Red Hook, NY: Curran Associates, Inc.)
Vickers, J. J., Grebel, E. K., & Huxor, A. P. 2012, AJ, 143, 86
Vickers, J. J., Li, Z.-Y., Smith, M. C., & Shen, J. 2021, ApJ, 912, 32
Xu, P., Zhu, X., & Clifton, D. A. 2022, CoRR, arXiv:2206.06488
Xue, X. X., Rix, H.-W., Zhao, G., et al. 2008, ApJ, 684, 1143
Xue, X.-X., Rix, H.-W., Yanny, B., et al. 2011, ApJ, 738, 79
York, D. G., Adelman, J., Anderson Jr., J. E., et al. 2000, AJ, 120, 1579