Identification of new M 31 star cluster candidates from PAndAS images using convolutional neural networks

Shoucheng Wang1,2,3, Bingqiu Chen2, Jun Ma1,3, Qian Long4, Haibo Yuan5, Dezi Liu2, Zhimin Zhou1, Wei Liu6, Jiamin Chen7, and Zizhao He3,8

1 Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, PR China
e-mail: majun@nao.cas.cn
2 South-Western Institute for Astronomy Research, Yunnan University, Kunming, Yunnan 650091, PR China
e-mail: bchen@ynu.edu.cn
3 School of Astronomy and Space Sciences, University of Chinese Academy of Sciences, Beijing 100049, PR China
e-mail: longqian@ynao.ac.cn
4 Yunnan Observatories, Chinese Academy of Sciences, Kunming 650216, PR China
5 Department of Astronomy, Beijing Normal University, Beijing 100875, PR China
6 National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China
7 School of Computer Science and Engineering, Central South University, Changsha 410083, PR China
8 Key Laboratory of Space Astronomy and Technology, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, PR China

Received 7 September 2021 / Accepted 12 November 2021

ABSTRACT

Context. Identification of new star cluster candidates in M 31 is fundamental for the study of the M 31 stellar cluster system. The machine-learning method convolutional neural network (CNN) is an efficient algorithm for searching for new M 31 star cluster candidates from tens of millions of images from wide-field photometric surveys.

Aims. We search for new M 31 cluster candidates from the high-quality g- and i-band images of 21 245 632 sources obtained from the Pan-Andromeda Archaeological Survey (PAndAS) through a CNN.

Methods. We collected confirmed M 31 clusters and noncluster objects from the literature as our training sample. Accurate double-channel CNNs were constructed and trained using the training samples. We applied the CNN classification models to the PAndAS g- and i-band images of over 21 million sources to search new M 31 cluster candidates. The CNN predictions were finally checked by five experienced human inspectors to obtain high-confidence M 31 star cluster candidates.

Results. After the inspection, we identified a catalogue of 117 new M 31 cluster candidates. Most of the new candidates are young clusters that are located in the M 31 disk. Their morphology, colours, and magnitudes are similar to those of the confirmed young disk clusters. We also identified eight globular cluster candidates that are located in the M 31 halo and exhibit features similar to those of confirmed halo globular clusters. The projected distances to the M 31 centre for three of them are larger than 100 kpc.

Key words. galaxies: star clusters: general – galaxies: star clusters: individual: M 31

1. Introduction

Star clusters are excellent tracers for the studies of the formation and evolution of their host galaxies. In past decades, many works have been carried out to search and identify new star clusters in our nearest archetypal spiral galaxy M 31. We only mention a few of these works. Barmby et al. (2000) presented a new catalogue of 435 clusters and cluster candidates in M 31 using photometric data observed by the 1.2m telescope of the Fred L. Whipple Observatory and spectroscopic data observed by the Keck LRIS spectrograph (Oke et al. 1995) and the Blue Channel spectrograph of the Multiple Mirror Telescope. Galleti et al. (2004) identified 693 clusters and cluster candidates in M 31 in the 2MASS database and presented the Revised Bologna Catalogue (RBC) of confirmed and candidate M 31 globular clusters. Huxor et al. (2004) discovered nine globular clusters in the outer halo of M 31 using images from an INT Wide Field Camera survey of the region. Based on the observation data by the imager/spectrograph DoLoRes at the 3.52 m TNG telescope and the low resolution spectrograph BFOSC mounted at the 1.52 m Cassini Telescope of the Loiano Observatory, Galleti et al. (2007) identified five genuine remote globular clusters according to the radial velocities from the spectra of these sources and their extended nature from ground-based optical images. Hodge et al. (2010) identified 77 new star clusters from the Hubble Space Telescope (HST) WFPC2 observations of active star formation regions of M 31. Based on the visual inspection of the image data from the Sloan Digital Sky Survey (SDSS; Stoughton et al. 2002), di Tullio Zinn & Zinn (2013) identified 18 new clusters in M 31 outer halo. Huxor et al. (2014) discovered 59 globular clusters and two candidates in the halo of M 31 via visual inspection of images from the Pan-Andromeda Archaeological Survey (PAndAS; McConnachie et al. 2009). Johnson et al. (2015) presented a catalogue of 2753 clusters in M 31 using the images
from the Panchromatic Hubble Andromeda Treasury (PHAT) survey. Chen et al. (2015) found 28 objects, 5 of which are bona fide and 23 of which are likely globular clusters based on radial velocities obtained with Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; Cui et al. 2012) spectra and visual examination of the SDSS images. Wang et al. (2021) classified 12 bona fide star clusters from the cluster candidates in the literature based on their radial velocities obtained with LAMOST spectra.

Most of these previous works identified new M 31 star clusters from the images via visual inspection by naked eye. The modern wide-field galaxy surveys have provided us tens of millions of source images. This makes the traditional visual inspection procedures tedious and very time consuming. Recently, one of the machine-learning methods, the convolutional neural network (CNN; Fukushima 1980; LeCun et al. 1998), has been adopted by astronomers to identify and classify different types of objects and obtain the physical properties of these objects from their images, as the CNNs are particularly suitable for image-recognition tasks. Petrillo et al. (2017) applied a morphological classification method based on CNN to recognize strong gravitational lenses from the images of the Kilo Degree Survey (KiDS; de Jong et al. 2017). Liu et al. (2019) proposed a supervised algorithm based on deep CNN to classify different spectral type of stars from the SDSS 1D stellar spectra. Bialopetravićius et al. (2019) developed a CNN-based algorithm to simultaneously derive ages, masses, and sizes of star clusters from multi-band images. Bialopetravićius & Narbutis (2020) detected 3380 clusters in M 83 from HST observations with a CNN trained on mock clusters. He et al. (2020) applied CNN to images from the KiDS Data Release 3 and identified 48 high-probability strong gravitational lens candidates. Based on a multi-scale CNN, Pérez et al. (2021) identified star clusters in the multi-colour images of nearby galaxies based on data from the Treasury Project LEGUS (Legacy ExtraGalactic Ultraviolet Survey; Calzetti et al. 2015).

The PAndAS survey, which has a coverage of >400 square degrees centred on the M 31 and M 33, providing us the most extensive panorama of M 31 with large projected galactocentric radii. The primary goal of PAndAS is to provide contiguous mapping of the resolved stellar content of the halo of M 31 out to approximately half of its halo virial radius (in projection). The coverage of its imaging reaches a projected distance of $R \sim 150 \text{ kpc}$ from the centre of M 31, which makes it possible to identify globular clusters in the remote halo of M 31. Based on the data from PAndAS, Huxor et al. (2014) identified 59 globular clusters and two candidates in the halo of M 31 primarily via visual inspection. Martin et al. (2016) presented a comprehensive analysis of the structural properties and luminosities of the 23 dwarf spheroidal galaxies that fall within the footprint of the PAndAS, which represent the large majority of known satellite dwarf galaxies of M 31.

In this work, we search for new star cluster candidates in M 31, including old globular clusters and young disk clusters, from multi-band images of tens of millions sources released by the PAndAS survey. We have constructed double-channel CNNs and trained the models with empirical samples selected from the literature. The CNN-predicted M 31 cluster candidates are finally confirmed by naked eye to obtain high-probability candidates.

This paper proceeds as follows. In Sect. 2 we introduce the data and the selection of the training samples. We describe our CNN models in Sect. 3 and our search procedure in Sect. 4. Finally, we present our results in Sect. 5 and summarize in Sect. 6.

## 2. Data

### 2.1. PAndAS images

The PAndAS survey started collecting data in 2003 and completed the collection in 2010. The survey was conducted in $g$ and $i$ bands using the Canada-France-Hawaii Telescope (CFHT) located on Maunakea, Hawaii. The observations were carried out with the MegaCam wide-field camera (Boulade et al. 2003), which contains 36 $2048 \times 4612$ pixel CCDs. The effective field of view of the camera is $0.96 \times 0.94 \text{ deg}^2$, with a pixel scale of $0.187''/\text{pixel}$. The values of the typical seeing of the PAndAS survey are $0.67''$ and $0.60''$ for the $g$ and $i$ bands (Huxor et al. 2014), respectively. With the excellent image quality, we are thus able to identify new star clusters in M 31. The images were processed with the Cambridge Astronomical Survey Unit (CASU), including image processing, calibration, and photometric measurements. We adopted the processed stacked $g$- and $i$-band images and the merged catalogue of all detected sources as accessed from the PAndAS VOspace.

### 2.2. Training samples

To train the CNNs, a large number of positive and negative sample images are required. We adopted empirical training samples. We first collected the confirmed clusters in M 31 as the positive training sample. Wang et al. (2021) collected 1233 confirmed clusters from previous works (Galleti et al. 2004; Johnson et al. 2012; di Tullio Zinn & Zinn 2013; Huxor et al. 2014; Chen et al. 2015). We complemented this catalogue with the “Just star clusters in M 31” catalogue contributed by Nelson Caldwell, which contains 1300 M 31 clusters from the literature (Caldwell et al. 2009, 2011; Caldwell & Romanowsky 2016). These two catalogues were merged and cross-matched with the PAndAS source catalogue, which yielded a sample of 1509 confirmed M 31 clusters that have been detected by PAndAS. The PAndAS $g$- and $i$-band images of these sources were then collected. We examined these images by naked eye and excluded star-like, overexposed, underexposed, or contaminated images. This selection led to a sample of 975 confirmed M 31 star clusters with high-quality PAndAS images. Example images for the sources we excluded are shown in Fig. 1. We excluded 81, 29, 75, and 349 star-like, overexposed, underexposed and contaminated sources, respectively. Most of the rejected objects are located in a crowded field, that is, the bulge or disk of M 31, where there are usually contaminated sources nearby. In addition, most of the brightest ($i < 14$ mag) and faintest sources ($i > 23$ mag) were also excluded. Within the magnitude range of $14 < i < 23$ mag, the colour and magnitude distributions of our remaining 975 sample clusters are similar to those of the total catalogue of 1509 clusters, which indicates that there is no selection bias for our positive training sample within the magnitude range of $14 < i < 23$ mag.

We cut the image stamps of the remaining 975 confirmed M 31 clusters as our positive sample. The size of the samples was $56 \times 56$ pixels, which is about $10.47 \times 10.47 \text{ arcsec}^2$. Because the positive sample was small, we augmented the data to increase the training sample artificially. We flipped the images horizontally or vertically, which yielded 2925 images. The images were then randomly rotated by 90, 180, or 270$^\circ$, which resulted in 11700 images. Finally, we randomly shifted the images with

1. [https://www.canfar.net/storage/list/PANDAS/PUBLIC](https://www.canfar.net/storage/list/PANDAS/PUBLIC)
2. [https://www.cfa.harvard.edu/oir/eg/m31clusters/m31clusters.html](https://www.cfa.harvard.edu/oir/eg/m31clusters/m31clusters.html)
Fig. 1. Example PAndAS images of M 31 confirmed star clusters that have been excluded from our positive sample in the current work. From left to right and top to bottom: examples of star-like, overexposed, underexposed, and contaminated clusters, respectively. Stamps are centred on the catalogued PAndAS positions and cover 56 × 56 pixels.

For the negative (non-M 31 star cluster objects) samples in our current work, we prepared three catalogues.

(1) Sources that belong to M 31, but are not star clusters, including stars, planetary nebulae (PNe), HII regions, and symbiotic stars in M 31 were included in the negative sample. For these sources, we adopted the catalogues “Stars in the M 31 catalog”, “PNe in M 31”, “HII regions in M 31” and “Symbiotic stars in M 31” contributed by Nelson Caldwell in the current work.

(2) Foreground Galactic sources were included in the negative sample. We collected the Gaia DR2 sources (Gaia Collaboration 2018) that are distributed within 2° to the M 31 centre (RA = 00°42′44″30, Dec = +41°16′09″0; Skrutskie et al. 2006) and have parallaxes higher than 0.01 mas, that is, distances from the Sun shorter than 100 kpc.

(3) Background quasars and galaxies were also included in the negative sample. We collected sources from LAMOST DR6 (Luo et al. 2015) that have radial velocities higher than 50 km s$^{-1}$, which also contain some Milky Way foreground stars.

The catalogues listed above were merged and cross-matched with the PAndAS source catalogue, which yielded a sample of 100,501 non-M 31 clusters as our negative sample. Similarly as for the positive sample, we cut the PAndAS images of these sources into 56 × 56 pixel samples for CNN training.

3. Convolutional neural networks

We used both the PAndAS $g$- and $i$-band images. Thus we constructed double-channel CNNs, which are able to make full use of the morphological information of images in $g$ and $i$ bands.

The CNN is a deep-learning algorithm especially for image and video recognitions. The CNN is able to extract the features of input images and convert them into lower dimensions without losing their characteristics. A typical CNN contains an input layer and a series of convolutional and pooling layers, which are followed by a full connection layer and an output layer. Table 1 shows our CNN architecture. The first layer is the input layer, which is a double-channel layer fed with image stamps with sizes of 56 × 56 × 2 pixels (for both the PAndAS $g$- and $i$-band images). Following the input layer, the second layer is a convolutional layer that aims to maintain the spatial continuity of the input images and extract their local features. The convolutional layer is also a double-channel layer and consists of 64 kernels. Each kernel has a size of 5 × 5. We used the rectified linear unit (ReLU; Nair & Hinton 2010) activation function of this convolutional layer, which helps increase the non-linear properties and speeds up the training process considerably. The third layer is a pooling layer that aims to reduce the dimensionality of the data and the parameters to be estimated. It provides translational invariance to the network. It also helps to control overfitting (Scherer et al. 2010). We used max pooling (Boureau et al. 2010) with a kernel size of 2 × 2, which takes the maximum value in a connected set of elements of the feature maps.

To extract enough features of the training images, we adopted three pairs of convolutional and max pooling layers in the network. The second pair of layers, that is, the fourth and the fifth layers, are similar to the second and third layers, respectively. The third pair (the sixth and seventh layers) are slightly different from the first two pairs. We adopted 128 kernels for the convolutional and the pooling layer. Two fully connection layers (the eighth and ninth layers) followed after the last pooling layer, which causes the nodes to connecting well to each other and
Table 1. Architecture of our CNN.

| No. | Layer     | Kernel size (units) | Features (units) | Activation function |
|-----|-----------|---------------------|------------------|--------------------|
| 1   | Input     | 56 × 56 × 2         | 2                | –                  |
| 2   | Conv$_1$  | 5 × 5               | 64               | ReLU               |
| 3   | MaxPooling$_1$ | 2 × 2        | 64               | –                  |
| 4   | Conv$_2$  | 5 × 5               | 64               | ReLU               |
| 5   | MaxPooling$_2$ | 2 × 2        | 64               | –                  |
| 6   | Conv$_3$  | 5 × 5               | 128              | ReLU               |
| 7   | MaxPooling$_3$ | 2 × 2        | 128              | –                  |
| 8   | FC        | –                   | 1024             | ReLU               |
| 9   | FC        | –                   | 1024             | ReLU               |
| 10  | Output    | 1 × 2               | –                | Softmax             |

This architecture accommodates all possible dependences for the entire architecture. Each layer has 1024 neural units. We used the dropout technique to prevent overfitting. Dropout randomly drops units from the neural network during the training process. We set the value of the dropout to 0.5. Finally, the tenth layer is the output layer, which is also a fully connected layer and was set as a softmax function,

$$P_i = \frac{e^i}{\sum_j e^j}.$$  \hspace{1cm} (1)

The output value ($e^i$) is interpreted into the probability ($P_i$) denoting that a source is an M31 star cluster or a non-M31 star cluster.

We used the binary cross-entropy as the loss function of the real and predicted classification to train the network. Similar to He et al. (2020), we minimised the cross-entropy by adopting an Adam optimiser (Kingma & Ba 2014) to optimise the network parameters, and especially to minimise the output error. The Adam optimisation algorithm is an extension of the stochastic gradient descent. It introduces a friction term that mitigates the gradient momentum in order to reach faster convergence. We used a batch size of 50 images and trained the network for 80 epochs. The learning rate was set as 0.001.

Fig. 2. $i$-band magnitude distribution of confirmed M31 clusters in our positive training sample.

4. Searching for new M31 clusters

4.1. Data pre-processing and model training

Our candidate clusters were selected from all objects detected by the PAndAS, including those classified as point sources (with a negative morphology flag), extended sources (with a positive morphology flag) and noise sources (with a morphology flag of zero) in the PAndAS source catalogue. This is because that there are confirmed M31 clusters from our positive sample in all the three categories. We selected object that have $13 < i < 23$ mag. This range in magnitude encompasses the range of the confirmed M31 clusters (see Fig. 2). This yielded 21 245 632 sources as our input sample.

We aim here to search for new M31 clusters from the combined $g$- and $i$-band image stamps of over 21 million sources from PAndAS by CNN. The input of machine-learning algorithms should be numerical. We therefore transferred the PAndAS $g$- and $i$- images into 56×56×2 Python numpy float arrays. To make the training procedure fast and efficient, the input images were then normalised by constraining the pixel values into a small and symmetric range. Similar to Liu et al. (2019), we adopted the standard z-score method, which is defined as

$$Z = \frac{M - \mu}{\sigma},$$  \hspace{1cm} (2)
Table 2. Catalogue of the 117 new M 31 star cluster candidates.

| Name          | RA (J2000)       | Dec (J2000) | g (mag) | i (mag) | R proj (kpc) | e     | PA (°) | Score |
|---------------|------------------|-------------|---------|---------|--------------|-------|--------|-------|
| Candidate1    | 0:15:05:956      | +36:27:38:25| 21.856  | 20.308  | 96.7         | 0.08  | 15.5   | 6     |
| Candidate2    | 0:30:32:016      | +37:20:41:34| 21.133  | 19.865  | 61.5         | 0.18  | 11.8   | 6     |
| Candidate3    | 0:34:42:035      | +39:31:24:42| 20.771  | 19.996  | 31.2         | 0.07  | −21.13 | 8     |
| Candidate4    | 0:37:16:566      | +40:06:36:16| 20.949  | 19.984  | 20.9         | 0.16  | 18.39  | 6     |
| Candidate5    | 0:37:48:369      | +39:44:29:92| 21.991  | 21.156  | 24.1         | 0.32  | −27.48 | 7     |

Notes. The full table is available at the CDS and also via the website http://paperdata.china-vo.org/Wang.SC/2021/table2.fits.

Fig. 5. PAndAS g- and i-band images of the six new M 31 cluster candidates that have the highest summed score from the inspectors. The coordinates of the objects are labelled.

where $M$ represents the flux value of the individual pixel, $\mu$ is the mean flux value of the image stamp, and $\sigma$ is the standard deviation. The standard $z$-score is a commonly adopted method in machine learning and is highly efficient. The standard $z$-score preprocessing was applied to our training sample and to the input sample.

In the beginning of the training, we labelled the positive and negative samples with the scheme called 'one hot coding' (Lantz 2015), which is the most widely used coding scheme to preprocess categorical attributes. We randomly divided the training sample (both positive and negative) into two parts, 80% of the sources to train the model, and 20% to test the training accuracy.

We applied our CNN to the training sample. To examine the accuracy of the training model, we adopted the F1 score (van Rijsbergen 1998) and confusion matrix, which are excellent indicators of the accuracy of binary classification models. An F1 score value of 1 denotes the best accuracy of training, and 0 represents the worst. It is defined as

$$ F_1 = 2 \cdot \frac{P \cdot R}{P + R}. $$

A51, page 5 of 9
where $P$ presents the precision and $R$ the recall. The precision $P$ and recall $R$ are defined as

$$P = \frac{TP}{TP + FP},$$

$$R = \frac{TP}{TP + FN},$$

where $TP$ means that positive classes are predicted as positive classes, while $FP$ means that negative classes are predicted as positive classes, and $FN$ denotes the false negative, which predicts the positive class as a negative class. We show in Fig. 3 the confusion matrix of our classification results. From the confusion matrix we can obtain that our CNN classification model has a precision = 0.996, a recall = 0.980, an F1 score = 0.988, and an accuracy = 0.988.

4.2. CNN prediction

The trained CNN classification model was applied to all 21 million sources in our input sample. To achieve high confidence, we confirmed an CNN experimental candidate with its prediction probability corresponding to a positive sample higher than 0.95. Our model predicted 65,548 candidates in all observational fields. We made a quick scan of the images of these candidates. These candidates are all extended sources. Point sources are well excluded by the CNN model. However, the size of the sample of the predicted possible candidates was still too large, and some elliptical galaxies contaminated the sample. The contaminations were mainly because the negative sample did not contain enough galaxies. To reduce the resulting sample size and exclude the extragalactic contamination, we processed a second CNN classification using the same network and the same positive training sample, but a different negative training sample. We cross-matched the 65,548 candidates predicted from the first CNN model with the SIMBAD database, which yielded 4363 sources classified as galaxies. The images of these galaxies were augmented in part with the same augmentation procedure as for the positive training sample (see Sect. 2.2). This yielded a negative training sample of 38,592 galaxies. Together with the previous positive sample, they were adopted to train the second CNN classification model. Again, 80% of the positive and negative training sample were adopted for training and 20% to test the accuracy. We obtained an F1 score = 0.980 and an accuracy = 0.981 for the second CNN network.

The second network was then applied to the 65,548 candidates predicted by the first CNN model. A prediction probability threshold of 0.98 was adopted to obtain the final CNN output, which yielded 5092 M 31 star cluster candidates. We cross-matched these candidates with the SIMBAD database and obtained 523 objects. Three hundred and eighty-four of the 523 objects are galaxies, a galaxy in a cluster of galaxies, an X-ray source, an infra-red source, a radio source, and so on. We visually inspected the PAndAS images of these 523 SIMBAD objects. Except for the confirmed clusters, none of the objects can be classified as bona fide clusters. They were thus excluded from our sample. We also cross-matched our candidates with catalogues from the literature (e.g. Narbutis et al. 2008) and found no counterparts.

5. Results and discussion

The 117 new M 31 cluster candidates are listed in Table 2. Each row of Table 2 contains the coordinates (RA and Dec), PAndAS $g$- and $i$-band magnitudes, projected distances to the centre of M 31 $R_{\text{proj}}$, ellipticity $e$, position angle (PA), and the summed score from the five inspectors of one identified M 31 cluster candidate. The PAndAS $g$- and $i$-band magnitudes were adopted from McConnachie et al. (2018). The values of $e$ and PA of the individual clusters were calculated from the PAndAS $i$-band image with the STSDAS ELLIPSE task (Science Software Branch of the Operations & Engineering Division at STScI 2012). The catalogue is available in electronic forms at the CDS. In Fig. 5 we show the PAndAS $g$- and $i$-band images with enhanced contrast of the six new M 31 cluster candidates that have the highest summed score (10) from all the inspectors. Their morphological appearances are similar to those of the confirmed M 31 clusters.

The spatial distribution of the new cluster candidates is shown in Fig. 6. Eight cluster candidates are located in the halo of M 31, with $R_{\text{proj}} > 25$ kpc. The farthest candidate is about 158 kpc away from the centre of M 31. Most of the new cluster candidates located in the disk of M 31 are located southeast of M 31. This is because the disk clusters we identified are mainly...
young clusters. The disk northwest of M 31 was observed by the PHAT project (Dalcanton et al. 2012); the young clusters were previously identified (Johnson et al. 2012, 2015).

5.1. Morphology

Here we confirm the morphological similarity between our newly identified M 31 cluster candidates and the confirmed clusters. In Fig. 7 we show the morphology of six example M 31 clusters from our positive training sample. The three example clusters plotted in the upper panel of Fig. 7 are young disk clusters selected from Johnson et al. (2012), and the other three are old halo clusters from Huxor et al. (2014). Overall, young disk clusters tend to be more symmetric and compact. In Fig. 8 we show the morphology of 12 example M 31 cluster candidates that we newly identified. Most of our candidates are located southeast of the M 31 disk, with projected radii $R_{\text{proj}} < 20$ kpc. We suggest that these candidates are young disk clusters. The first row of Fig. 8 shows four example clusters that are located in the disk. Their morphology is analogous to that of the young disk clusters from Johnson et al. (2012). We also identified eight new clusters in the halo of M 31 (with $R_{\text{proj}} > 25$ kpc). They are also shown in Fig. 8 and exhibit similar features as the globular clusters from Huxor et al. (2014).

5.2. Radial distribution

Figure 9 show the radial distribution of our newly identified 117 M 31 cluster candidates. Most of the clusters are located in the disk of M 31 and are concentrated in the area with galactocentric radii $R_{\text{proj}} < 10$ kpc. They are mainly young clusters with a morphology analogous to those from Johnson et al. (2012). Eight clusters are located in the halo of M 31. Three of them have projected radii larger than 100 kpc. They are all old clusters and have an extended radial profile, according to their morphology.

5.3. Colour-magnitude diagram

In Fig. 10 we plot the colour-magnitude diagram of all our newly identified M 31 cluster candidates. The young disk clusters from Johnson et al. (2012) and halo globular clusters from Huxor et al. (2014) are overplotted for comparison. We adopted the PAndAS $g$ and $i$-band magnitudes. Similar as in Wang et al. (2021), the reddening values of the individual objects were either adopted from the literature or the median reddening values of all the known clusters within 2 kpc. Significant discrepancies are visible between the young disk clusters from Johnson et al. (2012) and the halo globular clusters from Huxor et al. (2014). Young clusters are systematically bluer and fainter. Our newly identified M 31 clusters candidates have similar colours and magnitudes as the Johnson et al. young clusters, which suggests that most of them are young disk clusters.

6. Summary

We have constructed double-channel CNN classification models to search for new M 31 star clusters. Confirmed M 31 clusters and non-cluster objects from the literature were selected as our training sample. The trained CNN networks were applied to the $g$- and $i$-band images of over 21 million sources from PAndAS,
which have predicted 4569 machine-learning candidates. After human inspection, we identified 117 new high-probability M 31 cluster candidates. Most of our candidates are young clusters and are located in the disk of M 31. Eight candidates are located in the halo of M 31, with $R_{\text{proj}} > 25$ kpc. The farthest candidate has a projected distance of $R_{\text{proj}} = 158$ kpc. We also compared the morphology, colours, and magnitudes of our newly identified clusters with those of the young disk clusters from Johnson et al. (2012) and halo globular clusters from Huxor et al. (2014).
Obvious systematic discrepancies are visible between the young and old clusters. The features of our new candidates that are located in the disk are similar to those of the young clusters from Johnson et al. (2012), while the eight candidates located in the halo of M 31 exhibit similar features as the globular clusters from Huxor et al. (2014).

Contaminants such as M 31 HII regions, background galaxies, and foreground stars may exist in our catalogue. Spectroscopic observations of these candidates for a final identification in the future are required.

Acknowledgements. This work is partially supported by National Key R&D Program of China No. 2019YFA0405501 and 2019YFA0405503, National Natural Science Foundation of China (NSFC) No. 11803029, 11873053 and 11773074, and Yunnan University grant No. C619300A034. This work has made use of data products from the Pan-Andromeda Archaeological Survey (PAndAS). PandAS is a Canada-France-Hawaii Telescope (CFHT) large program that was allocated 226 hours of observing time on MegaCam. This work also has utilized the data (catalogues) from the Guoshoujing Telescope (the Large Sky Area Multi-Object Fibre Spectroscopic Telescope, LAMOST), Gaia Data Release 2 and websites by Nelson Caldwell.

References

Barmby, P., Huchra, J. P., Brodie, J. P., et al. 2000, AJ, 119, 727
Bialopetravičius, J., & Narbutis, D. 2020, AJ, 160, 264
Bialopetravičius, J., Narbutis, D., & Vansevičius, V. 2019, A&A, 621, A103
Boulade, O., Charlot, X., Abbon, P., et al. 2003, Proc. SPIE, 4841, 72
Bourey, Y. L., Bach, F., LeCun, Y., & Ponce, J. 2010, in 2010 IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 2559
Calderwood, N., & Romanowsky, A. J. 2016, ApJ, 824, 42
Calderwood, N., Harding, P., Morrison, H., et al. 2009, AJ, 137, 94
Calderwood, N., Schiavon, R., Morrison, H., Rose, J. A., & Harding, P. 2011, AJ, 141, 61
Calzetti, D., Lee, J. C., Sabbi, E., et al. 2015, AJ, 149, 51
Chen, B.-Q., Liu, X.-W., Xiang, M.-S., et al. 2015, Res. Astron. Astrophys., 15, 1392
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, Res. Astron. Astrophys., 12, 1197
Dalcanton, J. J., Williams, B. F., Lang, D., et al. 2012, ApJS, 200, 18
de Jong, J. T. A., Verdoes Kleijn, G. A., Eiben, T., et al. 2017, A&A, 604, A134
di Tullio Zinn, G., & Zinn, R. 2013, AJ, 145, 50
Fukushima, K. 1980, Biol. Cybern., 36, 193
Gaia Collaboration (Brown, A. G. A., et al.) 2018, A&A, 616, A1
Galleti, S., Federici, L., Bellazzini, M., et al. 2004, A&A, 416, 917
Galleti, S., Bellazzini, M., Federici, L., et al. 2007, A&A, 471, 127
He, Z., Er, X., Long, Q., et al. 2020, MNRAS, 497, 556
Hodge, P., Krienke, O. K., Bianchi, L., et al. 2010, PASP, 122, 745
Huxor, A., Tanvir, N. R., Irwin, M., et al. 2004, Satellites and Tidal Streams, 327, 118
Huxor, A. P., Mackey, A. D., Ferguson, A. M. N., et al. 2014, MNRAS, 442, 2165
Johnson, L. C., Seth, A. C., Dalcanton, J. J., et al. 2012, ApJ, 752, 95
Johnson, L. C., Seth, A. C., Dalcanton, J. J., et al. 2015, ApJ, 802, 127
Kingma, D. P., & Ba, J. 2014, ArXiv e-prints [arXiv:1412.6980]
Lantz, B. 2015, Machine Learning with R. Birmingham (UK: Packt Publishing)
LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proc. IEEE, 86, 2278
Liu, W., Zhu, M., Dai, C., et al. 2019, MNRAS, 483, 4774
Luo, A.-L., Zhao, Y.-H., Zhuo, G., et al. 2015, Res. Astron. Astrophys., 15, 1095
Martin, N. F., Ibata, R. A., Lewis, G. F., et al. 2016, ApJ, 833, 167
McConnachie, A. W., Irwin, M. J., Ibata, R. A., et al. 2009, Nature, 461, 66
McConnachie, A. W., Ibata, R., Martin, N., et al. 2018, ApJ, 868, 55
Narbutis, D., Vansevičius, V., Kodaira, K., et al. 2008, ApJS, 177, 174
Nair, V., & Hinton, G. 2010, Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair. Proceedings of ICML 27, 807
Oke, J. B., Cohen, J. G., Carr, M., et al. 1995, PASP, 107, 375
Pérez, G., Messa, M., Calzetti, D., et al. 2021, ApJ, 907, 100
Petriolo, C.E., Tortora, C., Chattjee, S., et al. 2017, MNRAS, 472, 1129
Scherer, D., Müller, A., & Behnke, S. 2010, Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition (Berlin: Heidelberg: Springer-Verlag), 92
Science Software Branch of the Operations & Engineering Division at STScI 2012, Astrophysics Source Code Library [record ascl:1206.003]
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Stoughton, C., Lupton, R. H., Bernardi, M., et al. 2002, AJ, 123, 485
van Rijsbergen, C. J. 1998, A Non-Classical Logic for Information Retrieval, Information Retrieval: Uncertainty and Logics (Heidelberg: Springer), 3
Wang, S., Chen, B., & Ma, J. 2021, A&A, 645, A115