Astronomy and Computing 34 (2021) 100442

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

Astronomical objects classification based on the Digitized First Byurakan Survey low-dispersion spectra

H. Astsatryan [a,*], G. Gevorgyan [a], A. Knyazyan [a], A. Mickaelian [b,*], G.A. Mikayelyan [b]

[a] Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia, 1, P. Sevak str., Yerevan 0014, Armenia

[b] V. Ambartsumian Byurakan Astrophysical Observatory of the National Academy of Sciences of the Republic of Armenia, Byurakan 0213, Aragatzotn province, Armenia

ARTICLE INFO

Article history:
Received 29 July 2020
Accepted 4 December 2020
Available online 16 December 2020

Keywords:
Astronomical data
DFBS
Machine learning
Data classification
Virtual Observatories
ArVO

ABSTRACT

The Digitized First Byurakan Survey is the largest and the first systematic objective-prism survey of the extragalactic sky. The detection, extraction, and classification of about 40 million spectra of about 20 million astronomical objects available in the survey require distinguishing the pixels containing photons from the source and the noise pixels per object. This paper aims at developing a service to classify the spectra of UV-excess galaxies, quasars, compact galaxies, and other objects in the survey. Supervised and unsupervised convolutional neural network deep learning algorithms have been developed and studied.

1. Introduction

The generated data volumes of astronomical sky surveys range from several terabytes to zettabytes (Scaife, 2020). Enormous data volume and complexity require developing and implementing innovative methods and novel approaches to data exploitation, such as Virtual Observatories (VO) (Hanisch, 2014). As a platform for launching astronomical investigations, VO provides access to massive data banks, software systems with user-friendly interfaces for data processing, analysis, and visualization, and even access to resources on which the work can be carried out. VOs enable astronomers, regardless of their location, to access the advanced computing facilities over the Internet. Constituted in 2002, the International Virtual Observatory Alliance (Quinn et al., 2004) brings together several national and international organizations, such as US Virtual Astronomical Observatory (Hanisch, 2012), German Astrophysical Virtual Observatory (Demleitner et al., 2007), or the European Virtual Observatory (Genova et al., 2015), which brings together many European countries.

The Armenian VO (ArVO) is a joint project between Byurakan Astrophysical Observatory and the Institute for Informatics and Automation Problems (Mickaelian et al., 2016) aiming at deploying an advanced virtual environment to meet data management challenges (Astsatryan et al., 2010). ArVO data resources' core is the Digitized First Byurakan Survey (DFBS) (Mickaelian et al., 2007), consisting of the extragalactic sky's largest prism survey. As the first systematic objective-prism survey of the extragalactic sky, DFBS covers 17,000 square degrees in the Northern sky and a high galactic latitude region in the Southern sky. Each DFBS plate contains low-dispersion spectra of about twenty thousand objects. The whole survey consists of about twenty million objects having several properties, like the color, broad emission or absorption lines, or spectral energy distribution (SED). Besides the DFBS, the datasets are obtained via 1 m, 0.5 m, and 0.2 m Schmidt (1.5 square degrees prism, photographic plates), 2.6 m (photographic plates and films) standard, and smaller old telescopes located at the Byurakan Astrophysical Observatory. These telescopes' metadata include names, coordinates, and magnitudes of the observed objects, equipment, receiver, emulsion, filters, date, time, and exposure of observations, sky and weather conditions, and observers. The DFBS datasets management system obtains the objects using different parameters (observing programs, telescopes, observing mode, dates, emulsions, or observers) (Mickaelian et al., 2009). The data is homogeneous in a unique survey with definite criteria, observing material, and methods. DFBS data is available in the open standard Flexible Image Transport System (FITS) digital format (Hanisch et al., 2001) to store, transmit, and process astronomical spectral objects (see Fig. 1).

FITS format allows identification of the field in astronomical coordinates and works with the available data. The DFBS contains many different spectral types depending on the object

* Corresponding authors.
E-mail addresses: hrach@sci.am (H. Astsatryan), aregmick@yahoo.com (A. Mickaelian).

https://doi.org/10.1016/j.ascom.2020.100442
2213-1337/

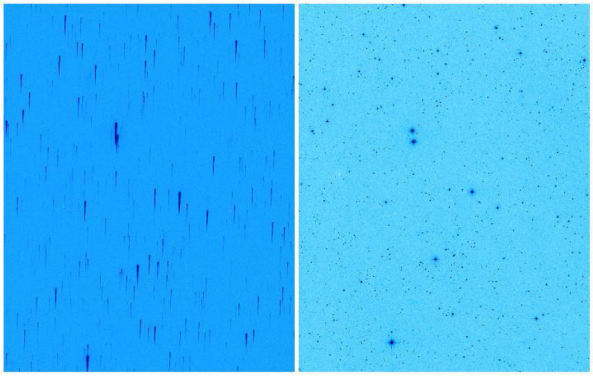

Fig. 1. DFBS spectral image and digitized sky survey direct image of the same area.

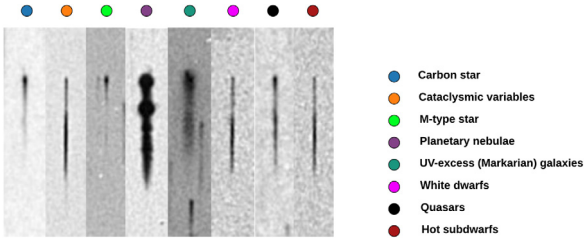

Fig. 2. DFBS spectra for various types of objects.

type, such as late-type stars, quasars, galaxies, or white dwarfs (see Fig. 2). The length, the shape, the spectral energy distribution, and available spectral lines allow identification of different types of objects.

The detection, extraction, and classification of spectra of about 20 million astronomical objects require distinguishing the pixels containing photons from the source and the noise pixels per object. The problem is more critical to identify very dim objects by their shapes and amount of brightness accurately, because of point-spread function convolution, noise, and blending. The DFBS contains up to 20 types (Massaro et al., 2008), but there is no spectral classification by their shapes. Machine learning (ML) paradigms are vital elements to extract and explore astronomical datasets enabling them to classify unexpected structures.

The study aims at developing a service based on convolutional neural network (CNN) to classify UV-excess galaxies, carbon stars, and other spectral objects available in the DFBS survey. As input data of CNNs, ASCII files of spectra and FITS images are used to classify the following objects' spectra:

- Ultraviolet-excess (UVX) galaxies (such as Markarian galaxies) – they have broader spectra than stars and longer UV (blue) part compared to other galaxies (Huchra, 1977). These objects appeared to contain exciting types of galaxies, such as Active Galactic Nuclei, including Seyferts, LINERs and some Quasars and Blazars, or Starburst galaxies. Very often, they are not distinguishable on low-dispersion spectra;
- Quasars (QSOs) – typically show blue spectra, have flat SEDs (spectral energy distributions) and strong/broad emission lines. Quasars are the most distant objects of the Universe and play crucial role in understanding the Cosmology;
- Compact galaxies (Seyferts, etc.) – often display stellar-like spectra, sometimes may display strong/broad emission lines;
- White dwarfs (WDs) – blue spectra, have broad absorption lines. WDs are very compact objects and are considered as the final stage of stellar evolution for most of the stars;
- Hot subdwarfs (sd) – very blue spectra, sometimes show broad absorption lines. Hot subdwarfs are important to understand the evolutionary transition between normal stars and WDs;
- Cataclysmic Variables (CVs) – blue spectra, in DFBS sometimes show emission lines. CVs are rather important for studies of close binary systems, stellar interactions and stellar evolution;
- Planetary Nebulae (PNe) – very strong emission lines and weak continuum. PNe are considered as slow mass ejection from central stars;
- Carbon (C) stars – extreme red spectra like short triangles, others (earlier subtypes) may display absorption bands. Carbon stars are important for understanding the chemical evolution of stars;
- Other late-type stars (such as M type) – red spectra with substantial red part and very faint blue part, absorption bands may be observed. M and other late-type stars are used to study the stellar evolution and Galactic kinematics.

2. Methodology

Both supervised and unsupervised learning methods have been developed to classify the extracted astronomical objects from the DFBS survey. As open-source platforms for machine learning, TensorFlow and Keras are used to train the network using the resources of the Armenian e-infrastructure (Astsatryan et al., 2015). Horizontal axis flipping, rotation, shifting, and noise injection augmentation techniques generate more training data from original data.

2.1. Data preprocessing

A three-step image processing algorithm has been developed to extract data from spectral images. In astrophysics, a threshold detection algorithm is quite popular to obtain the amount of light coming from each object to select pixels as sources or background. The image thresholding algorithm partitions an image into a foreground and background (Hajian et al., 2015). The threshold is applied for each spectral image to identify the objects from the background, and astronomical objects using the coordinates for two points (x1, y1, x2, y2) bounding rectangles that enclose them. It is assumed that all the pixels that appear different from the background correspond to astronomical objects. Before thresholding, an image, non-linear noise reduction Gaussian blur is used to blur the spectral images and remove noise (Buades et al., 2005). Then, the adaptive mean thresholding method is implemented to separate the foreground from the background. The threshold value is analyzed per each spectra in a window using its specific threshold value in the suggested adaptive thresholding method (see Fig. 3).

The astronomical coordinates (alpha and delta) to pixel coordinates (x and y) are converted for each astronomical object. For instance, if the right ascension and the declination are equal to 08h18m29.00s and 18d57m40.48s, then the object is taken from the plate FBS1477 using the plate fits file header information. Afterward, the astronomical coordinates to pixel coordinates are converted to get objects position in the spectral image. The x coordinate is equal to 4929, and y is equal to 4894 in the example, as mentioned above. The object's bounding rectangle is identified using the pixel coordinates of the object and thresholded image of the plate, which includes the object. Since the object's pixel coordinate is available, a boundary detection algorithm is applicable. The Theo Pavlidis algorithm has been implemented to find the object's bounding coordinates using simple tracing contour pixels based on a chain code to get the object's bounding

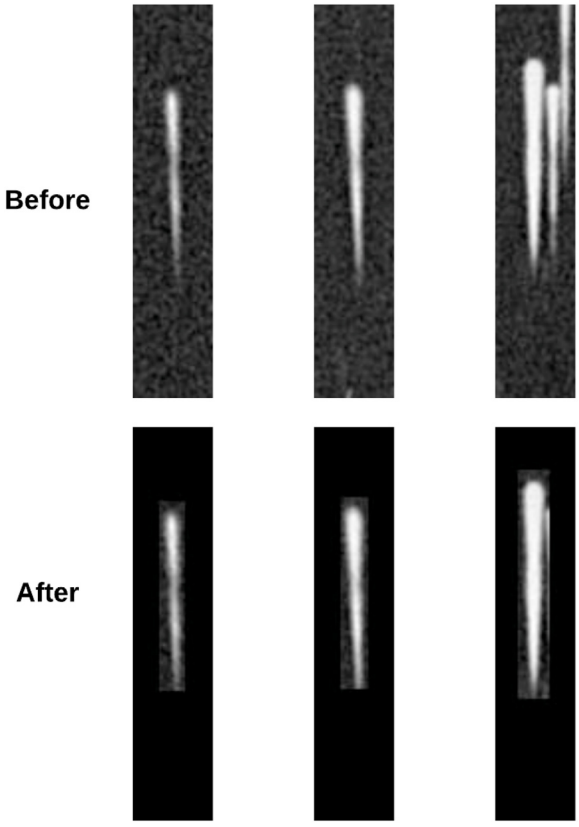

Fig. 3. The results of spectras bounding rectangle finding algorithm.

rectangle (Pavlidis, 2012). The algorithm considers only three adjacent pixels, e.g., front-left, front, and front-right. If all three pixels are white, the tracer turns right. It is possible to extract the object's 2D spectrum from the original plate image using the object's bounding rectangle (see Fig. 4).

2.2. Supervised learning

As a supervised learning approach, CNN deep learning algorithms (Indolia et al., 2018) have been developed and implemented on cleaned and ready data in the final stage. CNNs have deep feed-forward architecture and astonishing ability to generalize better than networks with fully connected layers. CNNs use the concept of weight sharing enabling to reduce the number of parameters to train the network. The limited number of parameters in CNNs overcome the suffer overfitting and train smoothly the datasets. The classification and feature extraction stages use for the learning process. An image to be classified is provided to the input layer, and the output is the predicted class label computed using extracted features from the image. A training dataset of example inputs and their corresponding desired outputs are used in a supervised learning system (see Fig. 5).

The spectral images represented as vectors are used as the input data for training and testing CNN models. Data normalizing scales data to fall within a smaller range, which helps speed up the training phase. The datasets are transformed into values between 0 and 1 by dividing the difference between actual and minimum values by the deviation of maximum and minimum values. Then the output is denormalized into the original data

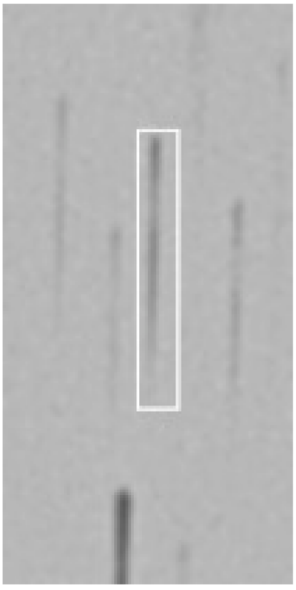

Fig. 4. An illustration of the extraction of an object's 2D spectrum from the original plate image.

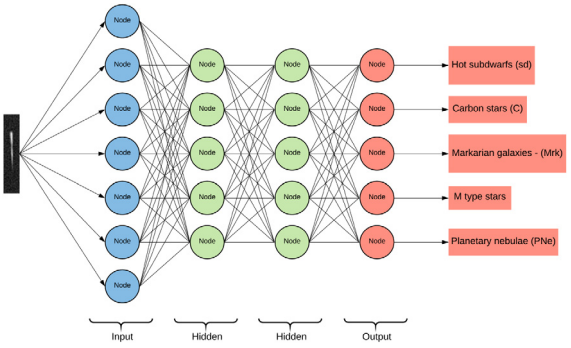

Fig. 5. Supervised learning scheme.

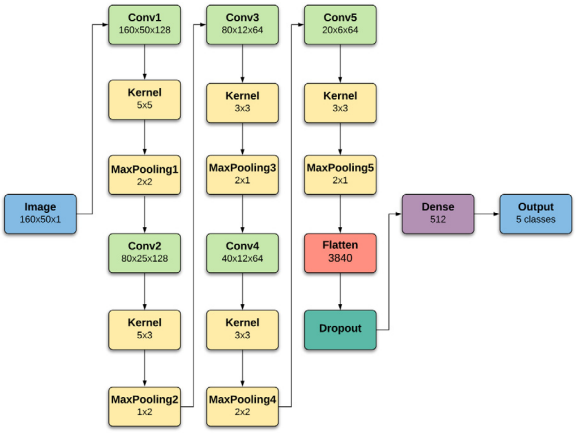

Fig. 6. CNN network architecture for image classification.

format for achieving the desired result. The network consists of 8 layers (see Fig. 6).

CNN accepts 160 × 50 pixels image as input, which is forwarded through the convolutional layer. As essential building

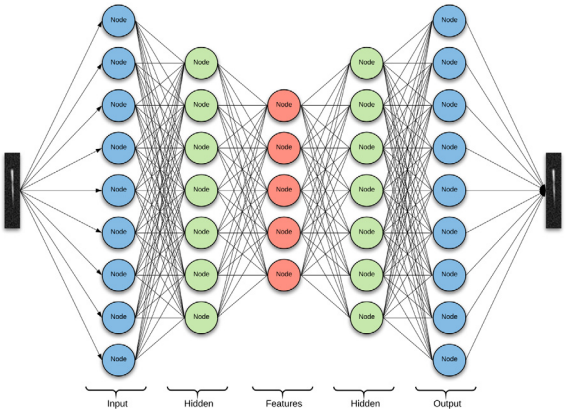

Fig. 7. Unsupervised learning scheme.

blocks for CNNs, the max pooling layers reduce the input dimensions and summarizes the most activated presence of a feature. The pooling function reduces map-size significantly and generates another output vector. Therefore, the pooling layers, which are implemented immediately after the convolutional layer, speed up the simulations and makes the detected features more robust. The convolution layer extracts the useful features from the input image with a filter. As an example, in the case of a 5X5 pixel filter, the Conv2D convolution layer computes the dot products between the image pixels' values and the weights defined in the filter. The final filter sizes are decided per each convolutional layer based on the experiments and the CNN results. A 2D convolution layer (Conv2D) means that the convolution operation's input is three-dimensional, while "2D convolution" refers to the movement of the filter in two dimensions. The flatten layer serves as a connection between the convolution and the output layers.

The ANN then evaluates the error according to some predefined cost function and computes appropriate corrections to the parameters. These prediction errors are propagated backward and use gradient descent to computation the parameter updates (Rumelhart et al., 1986). A rectified linear unit (ReLU) activation (output zero when the input less than zero and output equal to the input otherwise) is used for all hidden layers, as a functional mapping between inputs and outputs to learn and model the complex dataset (Nair and Hinton, 2010). The softmax activation function is used for the output layer. The model outputs are known classes of the spectrums.

2.3. Unsupervised learning

A training dataset of example inputs is implemented based on an unsupervised learning system consisting of input and hidden nodes (Kohonen, 1982). The network learns by associating several input pattern types with different hidden nodes (see Fig. 7).

The spectral images autoencoder represented as vectors are used as the input data. The autoencoder is implemented for an unsupervised ML model to decrease the shape of the input data. An autoencoder employs a symmetric structure composed of two main blocks:

- An encoder part that compresses the input into a low dimensional representation that contains the informative content of the data;
- A decoder part that is trained to reconstruct the input from the features extracted by the encoder.

Once the unsupervised pre-training is completed, the encoder part is thus a powerful automatic feature extractor that, completed with a suitable output layer, can be then fine-tuned in a supervised way to obtain the desired estimation. A small feature set for an object is created using the Autoencoder Artificial neural network based on convolutional layers, also called Convolutional Autoencoder. This artificial neural network helps to decrease the shape of the input data.

The density-based clustering algorithm has played a vital role in finding non-linear shapes based on clusterization density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is the most widely used density-based algorithm. It uses the concept of density reachability and density connectivity. It has two principal parameters sigma and minimum samples count. Since the features set are ready, the DBSCAN algorithm is used (Zhang, 2019) to classify the objects. The backpropagation algorithm is used for training and K-means for clustering to group similar data points together and discover underlying patterns (MacQueen et al., 1967). K-means looks a pre-defined number of clusters in a dataset.

3. Experimental results

The detection, extraction, and classification of about 20 million astronomical spectral objects require distinguishing the pixels containing photons from the source and the noise pixels per object. CNNs enable us to have accurate feature extraction and selection for star-galaxy classifiers by learning the local patterns. The results of a series of unsupervised CNN model experiments, using several models with different parameters and classes, are unsatisfactory due to the noisy astronomical data, as the overall accuracy is 38%, and the MSE error function value is 0.0012. Compere to unsupervised CNN models, the supervised models show high precision and recall, presented in the paper.

Based on a three-step image processing algorithm, the experiments show that supervised learning is a better approach for the studied datasets than unsupervised learning models. 78% of the dataset consists of 10465 images is used as training data, and 22% as testing data. The validation dataset is used to adjust and validate the model. The validation dataset is applied as a test dataset because of the limited size of the initial dataset.

Table 1 shows the initial and generated datasets, with the size of two and half million, to train and test the supervised CNN model and the classification accuracy per each object. The data is generated using the data augmentation techniques to increase the model's accuracy by decreasing both the training and validation losses. As the accuracy of train data less than the testing data, it prevents the model from overfitting. The model is not deep to be overfitted that easy. The dropout layer is added between existing layers to prevent overfitting by increasing the model accuracy as smoothly as possible. Also, a 50% high dropout coefficient is used, and the training process is stopped as soon as the validation loss rises.

These results are better for Markarian galaxies and planetary nebulae, while carbon stars' precision is quite low. The supervised learning model's[1] overall accuracy is about 87% achieved using different model configurations and several labeled datasets (see Fig. 8). Moreover, according to the figure, the learning curve shows a good feat of training and testing datasets.

As the loss function, the categorical cross-entropy has been used to train the network. In Fig. 9, the x-axis represents the number of iterations, while the y-axis represents the loss function value. The loss function value is decreased with an increase in the number of iterations and finally stabilized. Based on the experiments, the adadelta optimizer is used to adjust the network.

[1] The developed code used in this paper is available at: https://github.com/ArmenianVO/DFBSDataML.

Table 1
Supervised ML model accuracy and datasets.

Objects	Initial		Generated		Precision	Recall	F1-score
	Train	Test	Train	Test			
Hot subdwarfs (sd)	550	157	2750	785	95	96	96
Carbon stars (C)	331	94	1650	470	86	84	85
Ultraviolet-excess galaxies	305	86	1525	430	89	91	90
M type stars	104	44	770	220	80	49	61
Planetary nebulae (PNe)	11	4	131	48	100	100	100
Total	1301	385	6826	1953			

Table 2
CNN model classification results for two and a half million objects.

Accuracy	C Class	SD Class	PN Class	M Class	Mrk Class	Other
95%	403,556	154,623	6	6	29,336	1,929,075
90%	579,906	244,217	16	21	44,342	1,648,100
85%	697 535	322 233	278	62	61 782	1 434 712
80%	789 450	394 554	348	129	81 463	1 250 622

Table 3
CNN model classification results for about four million objects.

	C Class	SD Class	PN Class	M Class	Mrk Class	Other
95%	556,309	286,717	14	11	61,858	3,346,654
90%	829,577	440,737	38	46	91,670	2,889,531
85%	1,023,669	573,130	312	129	127,578	2,526,781
80%	1,181,754	695,987	392	265	167,559	2,205,606

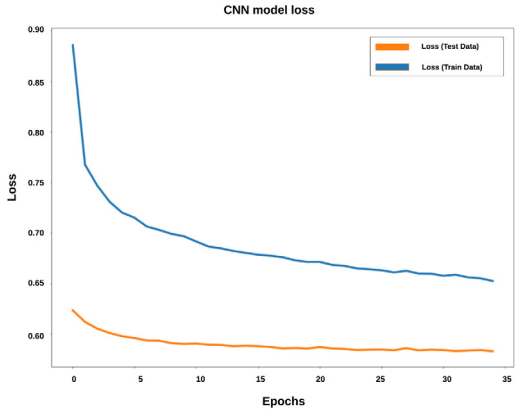

Fig. 8. CNN model accuracy.

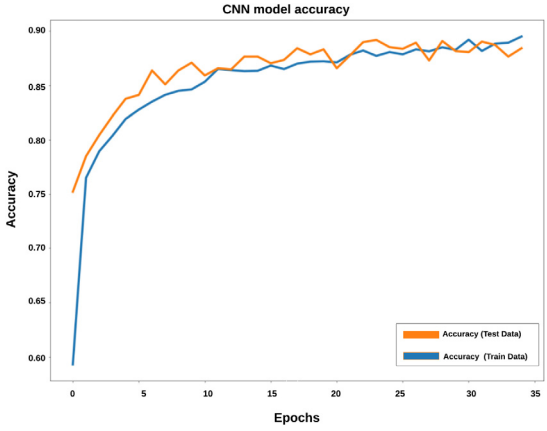

Fig. 9. Loss function of CNN training and testing.

The learning rate is 0.01. The experimental results are presented in Table 2 based on two and a half million objects, which is about 10% of the DFBS. The number of classified carbon stars, hot subdwarfs, and Markarian galaxies respectively are varied between 403556–789450 (16%–31%), 154623–394554 (7%–24%), 29336–81463 (1%–5%) when the accuracy decreased from 95% to 80%.

A variety of datasets sizes are carried out to tune the model and find data trends. Table 3 shows the results of about four million objects, where the number of classified carbon stars, hot subdwarfs, and Markarian galaxies respectively are varied between 556309–1181754 (13%–27%), 286717–695987 (7%–16%), 61858–167559 (1%–4%) when the accuracy decreased from 95% to 80%.

The linear regression analyzes show a high correlation between the total number and classified objects. For instance, in the case of sd and C classes, the linear regression correlation coefficient is equal to 98%–99%, and R^2 is equivalent to 96%–99%. According to the linear regression analyzes, it is assumed to expect to have at least six million carbon stars, three million hot subdwarfs, and one million Markarian galaxies in the DFBS survey, numbers that are significantly higher than expected before.

4. Conclusions and future work

In this paper, CNNs were introduced to classify UV-excess galaxies, quasars, compact galaxies, and other spectral objects in the DFBS survey. In the suggested supervised CNN model, the best results have been achieved in 34 epochs. The experiments show a good correspondence between the predicted and measured values, such as the overall accuracy is within 87%. Linear regression techniques have been implemented to forecast the number of objects in the DFBS survey expecting to have at least six million carbon stars, three million hot subdwarfs, and one million Markarian galaxies.

It is planned to increase the number of classes to predict and the accuracy. Based on the results, a cloud service will be deployed based on the suggested ML models.

CRediT authorship contribution statement

H. Astsatryan: Conception and design of study, Writing - original draft, Writing - review & editing. **G. Gevorgyan:** Data analysis and/or interpretation, Writing - original draft, Writing - review & editing. **A. Knyazyan:** Acquisition of data, Data analysis and/or interpretation, Writing - review & editing. **A. Mickaelian:** Conception and design of study, Writing - original draft, Writing - review & editing. **G.A. Mikayelyan:** Acquisition of data, Data analysis and/or interpretation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The paper is supported by the European Union's Horizon 2020 research infrastructures programme under grant agreement No 857645, project NI4OS Europe (National Initiatives for Open Science in Europe).

References

Astsatryan, H., Knyazyan, A., Mickaelian, A., Sargsyan, L., 2010. Web portal for the armenian virtual observatory based on armenian national grid infrastructure. Parallel Comput. Control Probl. 109–114.

Astsatryan, H., Sahakyan, V., Shoukourian, Y., Dongarra, J., Cros, P.-H., Dayde, M., Oster, P., 2015. Strengthening compute and data intensive capacities of armenia. In: 14th RoEduNet International Conference-Networking in Education and Research (RoEduNet NER). IEEE, pp. 28–33.

Buades, A., Coll, B., Morel, J.-M., 2005. A review of image denoising algorithms, with a new one. Multiscale Model. Simul. 4 (2), 490–530.

Demleitner, M., Gufler, B., Kim, J., Lemson, G., Nickelt-Czycykowski, I., Rauch, T., Stampa, U., Steinmetz, M., Voges, W., Wambsganss, J., 2007. The german astrophysical virtual observatory (gavo): archives and applications, status and services. Astron. Nachr. 328 (7), 713.

Genova, F., Allen, M.G., Arviset, C., Lawrence, A., Pasian, F., Solano, E., Wambsganss, J., 2015. Euro-vo—Coordination of virtual observatory activities in europe. Astron. Comput. 11, 181–189.

Hajian, A., Alvarez, M.A., Bond, J.R., 2015. Machine learning etudes in astrophysics: selection functions for mock cluster catalogs. J. Cosmol. Astropart. Phys. 2015 (01), 038.

Hanisch, R.J., 2012. Science initiatives of the US virtual astronomical observatory. Open Astron. 21 (3), 301–308.

Hanisch, R.J., 2014. The virtual observatory: I. Astron. Comput. 7, 1–2.

Hanisch, R.J., Farris, A., Greisen, E.W., Pence, W.D., Schlesinger, B.M., Teuben, P.J., Thompson, R.W., Warnock, A., 2001. Definition of the flexible image transport system (fits). Astron. Astrophys. 376 (1), 359–380.

Huchra, J.P., 1977. The nature of markarian galaxies. Astrophys. J. Suppl. Ser. 35, 171–195.

Indolia, S., Goswami, A.K., Mishra, S., Asopa, P., 2018. Conceptual understanding of convolutional neural network-a deep learning approach. Procedia Comput. Sci. 132, 679–688.

Kohonen, T., 1982. Analysis of a simple self-organizing process. Biol. Cybern. 44 (2), 135–140.

MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Oakland, CA, USA, pp. 281–297.

Massaro, E., Mickaelian, A., Nesci, R., Weedman, D., 2008. The digitized first byurakan survey. In: Aracne. p. 78.

Mickaelian, A., Astsatryan, H., Knyazyan, A., Magakian, T.Y., Mikayelyan, G., Erastova, L., Hovhannisyan, L., Sargsyan, L., Sinamyan, P., 2016. Ten years of the armenian virtual observatory. Astron. Soc. Pac. Conf. Ser. 505, 16–23.

Mickaelian, A., Nesci, R., Rossi, C., Weedman, D., Cirimele, G., Sargsyan, L., Gaudenzi, S., 2007. The digitized first byurakan survey — Dfbs. Astron. Astrophys. 464 (3), 1177–1180.

Mickaelian, A.M., Sargsyan, L.A., Astsatryan, H.V., Cirimele, G., Nesci, R., 2009. The dfbs spectroscopic database and the armenian virtual observatory. Data Sci. J. 0905280112.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. ICML.

Pavlidis, T., 2012. Algorithms for graphics and image processing. Springer Science & Business Media.

Quinn, P.J., Barnes, D.G., Csabai, I., Cui, C., Genova, F., Hanisch, B., Kembhavi, A., Kim, S.C., Lawrence, A., Malkov, O., et al., 2004. The international virtual observatory alliance: recent technical developments and the road ahead. In: Optimizing Scientific Return for Astronomy Through Information Technologies, Vol. 5493. International Society for Optics and Photonics pp. 137–145.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.

Scaife, A., 2020. Big telescope, big data: towards exascale with the square kilometre array. Phil. Trans. R. Soc. A 378 (2166), 20190060.

Zhang, M., 2019. Use density-based spatial clustering of applications with noise (dbscan) algorithm to identify galaxy cluster members. IOP Conf. Ser.: Earth Environ. Sci. 242 (4), 042033.