



Cloud-Based Machine Learning Service for Astronomical Sub-Object Classification: Case Study On the First Byurakan Survey Spectra

RESEARCH PAPER

HRACHYA ASTSATRYAN

STEPAN BABAYAN

AREG MICKAELIAN

GOR MIKAYELIAN

MARTIN ASTSATRYAN

*Author affiliations can be found in the back matter of this article

u[ubiquity press

ABSTRACT

The classification of astronomical objects in the Digitized First Byurakan Survey (DFBS), comprising low-dispersion spectra for approximately twenty million objects, presents challenges regarding performance and computational resources. However, considering the distinct spectral characteristics within subgroups, sub-object classification becomes crucial for a more detailed understanding of the dataset. The article addresses these challenges by proposing a comprehensive cloud-based service for classifying objects into spectral classes and subtypes, with a focus on carbon stars, white dwarfs / subdwarfs, and Markarian (UV-excess) galaxies, which are the primary objects in DFBS. By leveraging the power of cloud computing, it effectively handles the computational requirements associated with analyzing the extensive DFBS dataset. The service employs advanced machine learning algorithms trained on labeled data to classify objects into their respective spectral types and subtypes. The service can be accessed and utilized through a user-friendly interface, making it accessible to a wide range of users in the astronomical community.

CORRESPONDING AUTHOR:

Hrachya Astsatryan

Institute for Informatics and Automation Problems (IIAP),
P. Sevak 1, 0014 Yerevan,
Armenia

hrach@sci.am

KEYWORDS:

Telescopes; Surveys; Catalogs;
Byurakan Plate Archive;
Databases; Big Data; Data
Mining; Machine Learning;
Virtual Observatories

TO CITE THIS ARTICLE:

Astsatryan, H, Babayan, S,
Mickaelian, A, Mikayelian,
G and Astsatryan, M. 2024.
Cloud-Based Machine Learning
Service for Astronomical
Sub-Object Classification: Case
Study On the First Byurakan
Survey Spectra. *Data Science
Journal*, 23: 6, pp. 1–14. DOI:
[https://doi.org/10.5334/dsj-
2024-006](https://doi.org/10.5334/dsj-2024-006)

Astronomical research has produced vast datasets that challenge conventional data management and analysis techniques in recent decades, necessitating innovative approaches due to their sheer volume. Astronomical surveys systematically capture and catalog celestial entities and phenomena across the sky, serving as principal sources for new astronomical revelations. Nevertheless, their utilization remains limited, with only a handful of observatories and a fraction of astronomers depending on them, as pointed observations constitute the norm for most telescopes. A pioneering concept known as Virtual Observatories (VO) has emerged to address these challenges, offering astronomers enhanced efficiency in accessing, processing, and collaborating on astronomical data (Cui and Zhao 2007). VOs encompass interconnected networks of telescopes, data centers, and computational systems designed to support astronomers with interfaces for seamless data access and utilization of advanced computing resources.

The International Virtual Observatory Alliance (IVOA) was founded in 2002 to foster the coordinated and structured advancement of VOs (Quinn et al. 2004; Hanisch et al. 2015). IVOA, encompassing VO initiatives from 22 nations and 2 European projects, orchestrates the integration of astronomical resources and datasets by establishing technical standards for creating VOs and the seamless exchange of information between these platforms. Among these efforts, the Armenian Virtual Observatory (ArVO) has participated actively since 2005 (Mickaelian et al. 2016; Mickaelian et al. 2023).

ArVO is a project of the Byurakan Astrophysical Observatory (BAO) and the Institute for Informatics and Automation Problems, created to provide an advanced platform for astronomers to access and analyze astronomical data. BAO is famous for its many extragalactic surveys. The core of ArVO is the Digitized First Byurakan Survey (DFBS) (Mickaelian et al. 2007), which is the largest low-dispersion spectroscopic survey, containing about 40 million spectra for 20 million astronomical objects. The DFBS was carried out by renowned astrophysicist Benjamin Markarian and his colleagues during 1965–1980. It covers 17,000 deg² of the high galactic latitudes of the whole northern sky and part of the southern sky accessible from BAO. Before Gaia spectroscopy, it gave the largest number of spectra among all astronomical databases. These spectra provide spectral energy distribution (SED) and show several properties, like color, broad emission, or absorption lines. The DFBS contains spectra of many objects, such as late-type stars, quasars, galaxies, or white dwarfs. The length, shape, SED, and available spectral lines allow the identification of different types of objects. DFBS contains up to 20 distinguished types based on eye inspection and analysis, but their shapes need a proper spectral classification.

Nonetheless, analyzing these immense datasets presents formidable challenges owing to their voluminous nature and intricate underlying patterns. Automated extraction and classification mechanisms are necessary for studying such many objects. A model for the image-based classification of carbon stars, subdwarfs, Markarian galaxies, and other types of objects in the DFBS has been proposed by the authors (Astsatryan et al. 2021) to classify the DFBS objects based on their shapes and other visual characteristics. Utilizing convolutional neural networks combined with image preprocessing steps, the authors directly extracted and classified patches from the DFBS plates containing the desired objects. The evaluation showed hundreds of thousands of carbon stars, subdwarfs, and UV-excess (Markarian-type) galaxies in the DFBS, thus demonstrating the importance of employing a sub-object classification framework for further research. However, this approach has limitations. The current framework does not delve into sub-object classification, which restrains a more comprehensive understanding of the DFBS dataset. Extending the model to encompass sub-object categorization could usher a deeper insight into the dataset's diverse range of celestial bodies. Sub-classification enables researchers to identify commonalities and differences within a particular category to find correlations and potential relationships between various subgroups of objects. Furthermore, envisioning this model as a service would magnify its impact by granting researchers streamlined access to its capabilities, fostering collaboration, and expediting scientific advancements.

The paper aims to leverage the potential of cloud-based machine learning (ML) techniques to classify sub-objects to overcome the abovementioned limitations, enabling more efficient and accurate analysis. It presents a comprehensive end-to-end system designed to detect and extract celestial entities from astronomical plates and to categorize them into distinct subgroups, as illustrated in Figure 1.

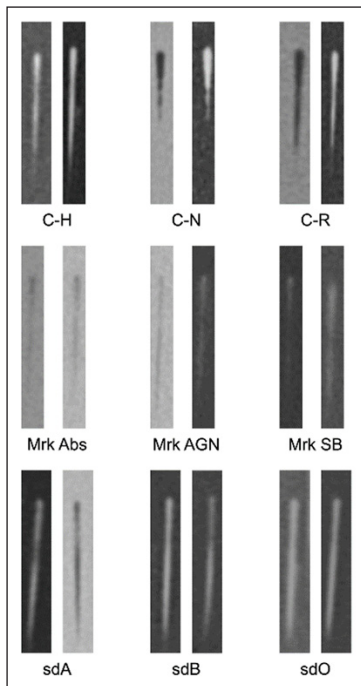


Figure 1 Examples of the subtypes of objects from the three most common object groups in the DFBS.

Within this framework, the service undertakes the classification of various object subtypes. The service classifies carbon stars as characterized by their carbon-rich atmospheres and exhibiting a distinct reddish hue due to lower temperatures. Simultaneously, the service addresses the classification of subdwarfs-metal-poor stars with a fainter and bluer profile than their counterparts, indicative of their early formation. In addition, the system tackles the classification of Markarian galaxies (Huchra 1977), notable for their distinctive attributes as active galaxies featuring intense star formation or active galactic nuclei (AGN), discernible through their strong UV excess.

In the sub-object classification, it is essential to differentiate the following spectral types:

- **Markarian galaxy**
 - Mrk AGN (Markarian Active Galactic Nuclei, 13^m – 16^m brightness)—active galactic nuclei within the sample. AGNs are characterized by a supermassive black hole at the center of a galaxy, which emits large amounts of radiation along the range of electromagnetic wavelengths due to the accretion of matter.
 - Mrk SB (Markarian Starburst Galaxies, 13^m – 16^m brightness)—starburst galaxies that undergo intense bursts of star formation, leading to high levels of stellar activity, emission lines, and young stellar populations.
 - Mrk Abs (Markarian Absorption-Line Galaxies, 13^m – 16^m brightness)—galaxies exhibiting prominent spectra absorption lines. Absorption-line galaxies are characterized by the vital contribution of stars that mostly have absorption lines, thus making up a broadened absorption-line spectrum of galaxies.
- **Carbon Stars**
 - C-R (Carbon Stars with Enhanced Red Emission, 13^m – 16^m)—exhibit enhanced red emission features in their spectra, attributed to the presence of dust grains and organic compounds in the circumstellar environment of these stars.
 - C-N (Carbon Stars with Enhanced Nitrogen, 13^m – 16^m)—exhibit enhanced nitrogen absorption features in their spectra. The enhanced nitrogen abundance results from nuclear processing and dredge-up of material within these stars.
 - C-H (Carbon Stars with Hydrogen, 13^m – 16^m)—exhibit the presence of hydrogen in their spectra. They are characterized by excessive carbon in their atmospheres, leading to distinctive molecular absorption features.
- **Subdwarf**
 - sdO (Subdwarf O Stars, 13^m – 16^m)—hot, luminous stars with helium-dominated atmospheres. They are more massive and hotter than subdwarf B stars. These are among the hottest known stars.

- sdB (Subdwarf B Stars, 13^m–16^m)—hot, compact stars with atmospheres primarily composed of hydrogen and helium. They are typically more massive than white dwarfs but less massive than main-sequence stars.
- sdA (Subdwarf A Stars, 13^m–16^m)—cooler and less massive than subdwarf B stars. They exhibit spectral features indicative of their hydrogen-dominated atmospheres.

By distinguishing these spectral types, the sub-object classification provides a finer level of categorization and enables more precise characterization and analysis of celestial objects within the mentioned groups. Besides, the photographic emulsion response is not linear (the characteristic curve characterizes it), causing additional difficulties both for the photometry and the classification. Thus, objects with different brightness have different shapes, and one should develop the classification for each magnitude range individually. For instance, very faint C stars are like dots, making distinguishing them in the DFBS fields difficult, as many defects and artifacts are also present. They resemble triangles when having 15^m–16^m, as only the red part is observable. For bright objects, one can follow some indefinite continuation of the spectra to the blue with any powerful red parts. To create a relevant classification scheme, we have grouped the spectra by magnitudes as follows: 13^m objects (12.6^m–13.5^m), 14^m (13.6^m–14.5^m), 15^m (14.6^m–15.5^m), 16^m (15.6^m–16.5^m), and 17^m (16.6^m–17.5^m). The last group is conditional, as the faintest spectra (~17^m near the survey limit) cannot be appropriately classified. On the other hand, bright objects (<12.5^m) are overexposed, and it is impossible to carry out any proper classification. So, relevant classification is mainly expected for the objects of 13^m, 14^m, 15^m, and 16^m groups.

The ML classification model contains a robust network to fit the data without underfitting, achieve high accuracy scores on relatively small data, and be scalable for inferring enormous amounts of data. Additionally, the paper aims to provide a comprehensive cloud-based service for the image-based classification of astronomical objects. The paper introduces the following key contributions:

- A robust convolutional neural network for sub-object classification of objects in DFBS is superior to the previous methods based on quantitative evaluation metrics on astronomical objects and sub-object classification datasets.
- A four-step classic image processing module to effectively identify and extract objects from DFBS ([Cloud-ML, 2021](#)), augmenting the previous image extraction pipeline and eliminating the necessity for resource-intensive object detection networks.
- Support for local and cloud-based experiments, accommodating diverse frameworks and adapting seamlessly to varying environments and data scales.
- A comprehensive cloud-based ML service encompassing training, visualization, testing, and fine-tuning components, all within a user-friendly interface tailored to the image-based classification of astronomical objects.

Despite the lower resolutions of the extracted objects, the proposed lightweight convolutional network and image extraction algorithm not only outperform the previous methods on group classification in DFBS but also attain satisfactory outcomes across both the training and validation sub-object classification datasets.

The rest of the paper is structured as follows. Section II reviews related work. The architecture and new concepts are presented in Section III. Section IV provides experimental results based on evaluations of different configurations of the testing dataset. Besides, the section describes the overall pipeline of the cloud-based infrastructure. Finally, conclusions and future research directions are discussed in Section V.

2 RELATED WORK

The field of astronomy is witnessing an unprecedented surge in both the volume and intricacy of data. Many projects explore and acquire spectral sky images ([Heymans et al. 2012](#); [York et al. 2000](#)). The scale of these endeavors has reached a point where manual classification is no longer feasible. ML algorithms have garnered significant attention in astrophysics, offering solutions to various challenges. However, the efficacy of these methods hinges on the availability of substantial datasets. Numerous ML and deep learning (DL) techniques have

been proposed for classifying astronomical objects, encompassing classical machine learning approaches, Convolutional Neural Networks (CNNs), and object detection networks. While many of these networks excel at classifying high-resolution objects and synthetic datasets, their performance needs to improve when confronted with low-resolution images and limited data volumes, often succumbing to overfitting.

CNNs have found widespread application in the realm of astronomical image classification. These networks operate by routing an image through convolutional layers, activating multiple neurons, with the number of active neurons aligning with the predetermined number of classes. For instance, Kim and Brunner (2016) employed CNNs for star-galaxy classification. Their network encompassed convolutional and pooling layers, culminating in two dense layers. The authors used a similar method with data preprocessing to classify Markarian galaxies, quasars, compact galaxies, and other objects in the DFBS.

As the data volume expands, the network architecture depth must be correspondingly enhanced. Residual connections avoid vanishing gradient problems and other training difficulties in deep networks (He et al. 2016). This concept has been leveraged by (Alawi and Al-Roainy 2021) to deploy and assess four distinct ResNet architectures to classify stars and galaxies. Their findings underscored the positive correlation between increased residual blocks and enhanced accuracy, mainly when ample data is available. Ethiraj and Bolla (2021) took a different approach by employing transfer learning networks, including Resnet50, DenseNet121, and Xception, supplemented with ImageNet weights, for their classification endeavors. Fine-tuning different layers within these networks demonstrated that DenseNet121 and Xception achieved the highest accuracy among the seven transfer learning networks assessed on the SDSS-IV dataset (Abolfathi et al. 2018). Nonetheless, it's essential to exercise caution when deploying large CNNs with comparatively smaller training datasets than those employed in the models above. Such networks are prone to overfitting due to the limited data, and their performance diminishes as regularizations increase, leading to slower learning rates.

Harnessing their aptitude for detecting and categorizing numerous objects, object detection neural networks are a unifying solution for addressing extraction and classification challenges within a singular framework. For instance, the versatility of modified Faster R-CNN has been harnessed to tackle detection and classification tasks (Jia et al. 2020). In a related context, Burke et al. (Burke et al. 2019) successfully employed Mask R-CNN to deblend and detect sources within multiband astronomical images and classify them. Notably, Mask R-CNN represents an advancement over Faster R-CNN, exhibiting favorable performance metrics at moderate intersection over union thresholds (He et al. 2017). Designed initially for RGB images, adapting Mask R-CNN to detect single-channel images necessitates code-level and training optimization-level adjustments. R-CNNs, two-stage detection models, exhibit significant drawbacks regarding efficiency and suitability when handling large quantities of images, rendering them an unsuitable option for rapid detection in extensive-scale scenarios.

Better solutions for detection in terms of time effectiveness are real-time networks. Thus, González et al. (2018) have used YOLO (You Only Look Once), a single-phase detector that works faster than the models described above. Currently, there are newer versions of YOLO that outperform the latter both in accuracy and speed. YOLO is speedy and learns generalizable representations of objects so that when trained on natural images and tested on the artwork, the algorithm outperforms other top detection methods. Although one-stage and two-stage object detection networks are end-to-end, they bring up some difficulties. First, data annotation must be performed, including classifying objects and marking their bounding boxes. Besides, detection networks are inferior to classification networks in terms of time complexity.

Authors have dedicated considerable efforts to harnessing ML techniques within the context of DFBS, yielding supervised and unsupervised approaches for object classification. Notably, these endeavors encompass detecting bounding boxes of objects on astronomical plates and classifying prominent astronomical entities, such as Markarian galaxies, Planetary Nebulae, and Carbon stars, employing supervised learning methodologies. These studies yielded an average accuracy of 87%, substantiating their efficacy. Moreover, the trained networks were extrapolated to predict the classification of other objects within the survey. This article stands poised to transcend the limitations observed in prior investigations and make substantial contributions to the advancement of astronomical data analysis and the pursuit of discoveries within this field.

3 METHODOLOGY

The astronomical object extraction and image-based classification methodology involves several main actions illustrated in Figure 2. The initial stage encompasses a four-step algorithm (depicted in Figure 2(a)) diligently orchestrated to extract objects from the DFBS plates. This multifaceted process is followed by data cleaning, normalizing, and transformation to establish harmonious compatibility with the ensuing ML algorithms. Afterward, the extracted objects are augmented and fed into a classification network for training and validation (Figure 2(b)). Finally, new and unlabeled objects are predicted using the previously trained network.

3.1 OBJECT EXTRACTION

A single astronomical plate in the DFBS (covering 16 square degrees of the sky) can contain about 10,000 to 25,000 objects, depending on how crowded the field is. On average, the spectral types of about 40 can be distinguished. Two novel approaches are presented: a four-

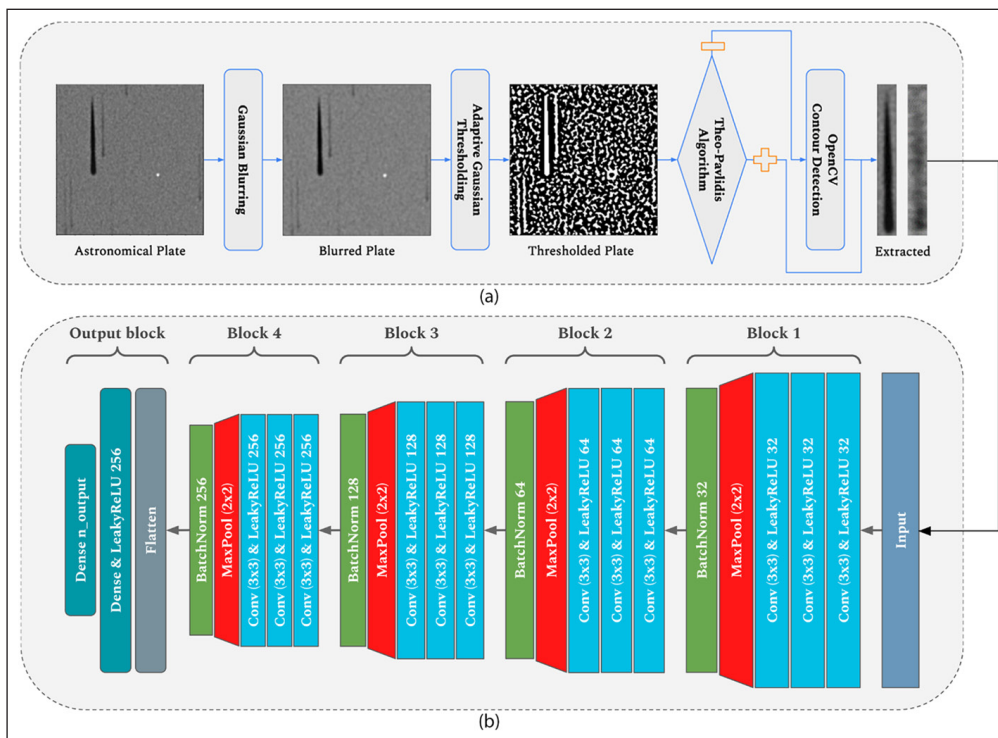


Figure 2 The workflow chart: (a) four-step image processing, (b) the proposed CNN architecture.

step algorithm to extract the labeled objects and an automated three-step algorithm for all the unlabeled objects within a specific filter range. The base of both algorithms is the three-step data preprocessing mechanism (Astsatryan et al. 2021), which can extract astronomical objects from grayscale images.

The three-step algorithm first blurs the plates by applying a Gaussian filter. Then, the grayscale plates are thresholded using the adaptive mean thresholding method, which sets the pixels' value to the maximum possible if the values are greater than the calculated threshold in their neighborhood and to the minimum otherwise. Therefore, the thresholding algorithm separates foreground objects and background noise. The third step involves marking the bounding boxes of the objects using Pavlidis' contour tracing algorithm (Pavlidis 1980). The algorithm takes a binary image and a starting point as input and returns the coordinates of the contour obtained by recursively tracing all the neighbors of the starting point of the same color. The initial points are calculated in the framework using the header file information of the plates, right ascension (RA), and declination (Dec) of labeled astronomical objects.

However, thresholding algorithms sometimes fail to distinguish between foreground and background. In poorly separated areas, foreground astronomical objects may have background fragments, detaching their different regions. In such cases, Pavlidis' contour tracing algorithm detects only the part of the object where the starting point is located.

The proposed approach presents the following improvements to overcome the abovementioned problems. First, adaptive Gaussian thresholding has been implemented instead of adaptive mean thresholding. Although the Gaussian approach is computationally expensive, it is more robust to noise and bright background artifacts than the mean thresholding. As a result, thresholded astronomical objects are clearly separated from the background and more objects are extracted. Besides, the preprocessing steps for object extraction are implemented separately from classification; they are applied to a fixed number of astronomical plates and, therefore, do not influence the processing time based on the size of the datasets. Furthermore, a constraint on the minimum height and width of the extracted image is introduced. As the experiments have shown, objects with a length of no more than 20 pixels are either background fragments or objects indistinguishable from the background. Added constraints filter unwanted objects automatically and improve overall accuracy. Finally, another preprocessing step as a backup method is added if the extracted image is invalid against the constraints. This step cuts the region around the starting point from the whole picture and searches for all the contours containing the starting point using a standard contour detection algorithm (Arbelaez et al. 2010). After filtering them, it selects the largest contour meeting the upper and lower size constraints.

The first two preprocessing steps are the same for extracting unlabeled objects. The object extraction step is implemented for two cases. In the first case, an inference dataset containing the name, RA, Dec, and other information of objects is provided. The framework uses the default settings by employing the four-step preprocessing algorithm for object extraction and feeding them into the classification network. Since different plates can contain the same objects and some objects may be extracted multiple times, the most frequently predicted class for these objects is returned as the classification output to consolidate the predictions of the network. In the second case, contours across all plates are searched, and those that meet the abovementioned constraints are extracted. This step can only be utilized for research purposes on the DFBS as the names and other identifiers of extracted objects are unknown; therefore, their classification can only provide general content information about astronomical plates.

3.2 SUPERVISED LEARNING

The extracted objects are systematically preserved as images, whereas their metadata, such as name, actual path, and parent plate affiliations, are stored in DataFrames. Before being fed into CNN, these images are accessed via their metadata and subjected to augmentation and normalization. Augmentation strategies encompass horizontal flipping, adjustments in brightness/contrast, random shifting, zooming, and rotation. This augmentation methodology effectively fortifies the model's robustness and generalization potential. After the transformation steps described above, the extracted objects are passed through a CNN for classification (see Figure 2(b)). The network takes 160×50 pixels size images (corresponding to the size of DFBS spectra) as input and returns n neurons as output, where n is the number of spectral types to be classified. The overall architecture consists of four blocks of convolutional, batch normalization, and pooling layers for feature extraction and dense layers for classification at the end. Each block integrates three consecutive convolutional layers, followed by max pooling and batch normalization layers. The first two blocks incorporate dropout layers at their conclusion, strategically introduced to curtail overfitting. The number of kernels of convolutions in consecutive blocks is 32, 64, 128, and 256, respectively, and the drop rates of dropout layers are $2/10$. The convolutional layers within each block operate with kernel sizes 3×3 , while all max pooling layers adhere to a 2×2 kernel size. The penultimate dense layer accommodates 256 neurons, adept at adapting training data without courting overfitting. Additional dropout layers were added after this layer and the fourth convolution block with $1/2$ and $3/10$ high dropout rates to prevent overfitting. All activation functions are set to LeakyReLU, as it avoids the 'dying ReLU' problem, which causes some neurons to always output zero values. Additionally, LeakyReLU preserves negative information and gives more flexibility to the network. The widely used softmax activation is removed from the last layer as it requires computationally expensive normalization to provide a probability distribution. This way, the network outputs logits, and a custom softmax function can be called on them for specific cases.

The training process of the proposed network on the sub-objects dataset was not straightforward due to the following reasons:

- **Data scarcity:** Usually, neural networks have an enormous number of parameters, and when the training dataset is small, they easily overfit. This prevents networks from generalizing and results in relatively lower performances on unseen datasets.
- **Dataset imbalance:** Most spectral classes have relatively small numbers of examples, and they often get confused by the network with other subclasses of the same group, which are visually similar and have dominant numbers of samples.
- **The low brightness and low contrast of the objects:** Although some spectral groups (e.g., Markarian galaxies) are essential for the research, they are faint. Thus, their samples are hardly distinguishable from the background, and their shapes are changed and distorted.

The network architecture, optimizer, loss function, and other hyperparameters were designed based on theoretical and empirically established assumptions to tackle these challenges. First, focal loss was employed, focusing on misclassified examples, giving them more weight on overall loss. This loss is especially effective on highly imbalanced datasets and has shown superiority over the commonly used cross-entropy loss in such cases (Tran et al. 2019). The focal loss adds two additional terms to standard cross-entropy. The first term is the focusing parameter, denoted by γ , which increases the weights of misclassified examples and down-weights well-classified ones. The second term, α , is added in practice to adjust the contribution of each class to the overall loss. Moreover, the drop rates of corresponding layers were chosen to balance the trade-off between overfitting and over-regularization of the model. The latter limits the model's capabilities and performs poorly on all datasets.

As one of the most popular gradient descent optimization algorithms, the authors leveraged the Adam optimizer (Bock and Weiß, 2019) for network optimization with a small learning rate. It combines an adaptive learning rate with the momentum-based approach to prevent overfitting and avoid local minimums by making more extensive updates where needed. The momentum-based approach speeds up the training process by adding the previous gradients to the current one with some factor and pushing the optimization process in the right direction. Besides, the adaptive learning approach assigns larger learning rates to small gradients and smaller learning rates to large gradients. This results in avoiding local minimums that cause small gradients and avoiding deviations affected by outliers. As a result, the training converges faster than with most other optimization algorithms and reaches higher scores (Kingma and Ba 2014). Finally, the authors employed the self-transfer learning approach to address the challenge of limited data by training the network on datasets featuring fewer classes (Pan and Yang 2009). This technique enables the network to leverage knowledge gained from these datasets and effectively overcome the issue of insufficient data. In this case, the network was pre-trained on the whole training dataset and then fine-tuned only on the subset of target subtypes. Performance gains were observed as the network learned the shape and properties of astronomical objects from a large amount of data during pre-training and learned only the given classification task on a small amount of data.

3.3 CLOUD-BASED SERVICE PLATFORM

Automated service provisioning is essential for big-scale image-processing tasks in astronomy. The suggested cloud-based ML service for the image-based classification of celestial bodies and beyond is based on Google Colaboratory, a cloud service providing Jupyter Notebook environments for educational and research purposes (Carneiro et al. 2018). The choice of Google Colaboratory is based on its multiple advantages over other alternatives. It is within reach for everyone, is easy to use, and provides comprehensive environments with pre-installed packages for running Jupyter notebooks. Moreover, the resources provided by Colab include free GPUs and TPUs that are essential when processing immense volumes of astronomical data.

The service comprises Colab notebooks for neural network training, fine-tuning, testing, and visualizations. It aims to provide astronomers with an environment to analyze astronomical data without knowledge of ML techniques. The flowchart of the proposed service is depicted in Figure 3. It consists of six general blocks, some of which differ based on the type of the selected processing pipeline.

The notebooks for training and evaluation are stored in the provided GitHub repository. Initially, the notebooks clone the code repository into the Colab environment. Next, the environment

variables, such as paths, directories, and processing devices, are chosen. Data is assembled after completing the environment configuration, and datasets are loaded. To process their data, users can adjust hyperparameters, such as input and output sizes, the number of elements in

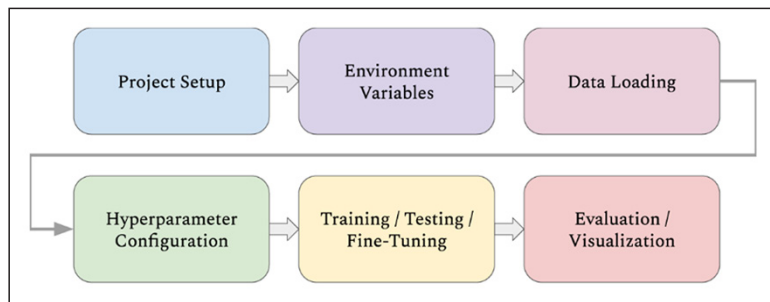


Figure 3 The flowchart of the cloud-based service.

training and testing batches, and checkpoints, in case they want to continue their unfinished pipelines. Several neural network architectures, such as ResNet and MobileNetV2 (Sandler et al. 2018), are also integrated into the service and can be employed by simply changing one line of code. These networks are already configured with tuned hyperparameters for classification tasks of close domains. The model checkpoints, trained on the presented datasets, are available for testing and fine-tuning.

Two evaluation types are provided to study the performance of different pipelines: step-by-step reports containing various classification metrics and TensorBoard-based monitoring measures. The former includes assessment parameters, such as precision, recall, and accuracy, and the latter consists of training accuracy and loss plots.

4 EXPERIMENTAL RESULTS

The section provides a comprehensive performance evaluation, including details about the sub-group and group classification datasets, optimal hyperparameters, and training configurations, like the number of epochs, batch size, or learning rate schedule. Then, the quantitative results of the proposed framework are presented on both datasets, focusing on classification performance evaluation metrics and other quantitative scores. These results provide insights into the accuracy of classifying astronomical objects. They are also compared with the previous work to demonstrate the superiority of the proposed method on both tasks. The trained model was inferred on two massive datasets from the DFBS by employing the second object extraction approach for unlabeled objects discussed in Section 3.1. The results are revealed later in this section.

The astronomical objects have been extracted from astronomical plates based on the pipeline introduced in the previous work. The one key difference is the four-step preprocessing mechanism described in Section 3.2 instead of the preceding three-step approach. The enhanced method better outlines astronomical objects and automates the invalid objects filtering process. As a result, it extracts up to 30% more objects than the previous method, given the same dataset containing objects' RA and Dec parameters.

The sub-object classification dataset consists of 2,107 training and 405 validation samples. These numbers could be higher because some classes only have 40–50 samples. The similarity of image characteristics of sub-objects within the same group (e.g., sdB and sdO subdwarfs) combined with one class having relatively fewer samples than the other class results in confusing the model always to predict the dominant class. Experiments have been conducted on three sub-groups taken from the extracted objects to avoid biased framework results caused by low numbers of training samples in some of the classes.

The comparison of the frameworks on groups' classification has been conducted by utilizing the publicly available dataset of the previous method in the GitHub repository (Astsatryan et al. 2021). The dataset has been divided into training and validation splits with 1,478 and 258 observations, respectively. The training pipelines have been reproduced for both methods on

4.1 EXPERIMENTAL SETTINGS

Training the developed model on all datasets was performed using the same Adam optimizer with a learning rate of $5e^{-4}$. Batch sizes for training sets were set to 256, with sizing for validation sets at 128. A learning rate scheduler, which reduced the learning rate by 0.9 factor after every five epochs, was employed to ensure stable training during higher steps. α and γ parameters of the loss function were set to 0.25 and 3, respectively. The augmentation steps described in Section II were applied to the data on both datasets during training.

The desirable results, demonstrated in Tables 1 and 2, were achieved in 126 and 131 epochs for the introduced framework, while after 128 and 137 epochs of training the former model. The tables represent the scores of two models, divided by a forward slash to compare their relative performance, where the proposed method is on the right. Figures 4 and 5 display training and testing accuracy curves for corresponding datasets. They also demonstrate the effectiveness of utilizing the self-transfer learning approach. Notably, accuracy metrics are high during the first epochs, and the model converges faster and reaches high accuracies in fewer epochs. The epochs for which checkpoints were selected and saved as results were determined by considering relevant classification metrics and accuracy plots to prevent picking overfitted models. Under the PyTorch framework, the single-node training pipeline and experiments were implemented on NVIDIA GeForce RTX 3080 GPU and Intel Core i9-11900 CPU within the research cloud infrastructure (Astsatryan et al. 2004).

4.2 QUANTITATIVE RESULTS

Four representative evaluation metrics assess the performance of the classification network:

	SUPPORT	PRECISION	RECALL	F1-SCORE
C-H	626 + 117	0.89 / 0.93	0.91 / 0.97	0.90 / 0.95
Mrk SB	664 + 132	0.94 / 0.95	0.89 / 0.95	0.91 / 0.95
sdB	817 + 156	0.94 / 0.98	0.96 / 0.96	0.95 / 0.97
Accuracy	2107 + 405			0.93 / 0.96
Macro avg	2107 + 405	0.92 / 0.96	0.92 / 0.96	0.92 / 0.96
Weighted avg	2107 + 405	0.93 / 0.96	0.93 / 0.96	0.93 / 0.96

	SUPPORT	PRECISION	RECALL	F1-SCORE
C	362 + 63	0.82 / 0.87	0.89 / 0.87	0.85 / 0.87
M	169 + 29	0.70 / 0.74	0.48 / 0.69	0.57 / 0.71
Mrk	333 + 58	0.91 / 0.94	0.88 / 1.00	0.89 / 0.97
PN	13 + 2	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
sd	601 + 106	0.93 / 1.00	0.98 / 0.98	0.95 / 0.99
Accuracy	1478 + 258			0.88 / 0.93
Macro avg	1478 + 258	0.87 / 0.91	0.85 / 0.91	0.86 / 0.91
Weighted avg	1478 + 258	0.87 / 0.93	0.88 / 0.93	0.87 / 0.93

Table 1 Comparison of classification reports for two frameworks proposed by the authors on the sub-objects' dataset. Support denotes the number of samples in the corresponding class (train + test), blue color denotes the better score.

Table 2 Comparison of classification reports for two frameworks proposed by the authors on the group classification dataset.

precision, recall, and f1-score for each class and accuracy for the overall classification. Precision is the ratio of the correctly predicted positives to all predicted positives for a given class. In contrast, recall is the ratio of the correctly predicted positives to all ground-truth positives. The f1-score is the harmonic mean of precision and recall, placing more weight on the lower value

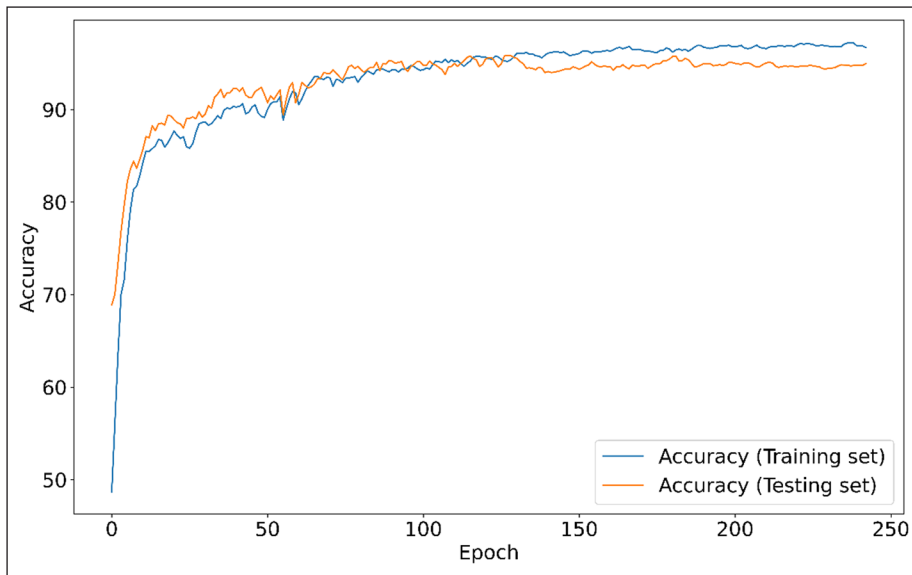


Figure 4 The accuracy of training and testing sets for the sub-object classification dataset.

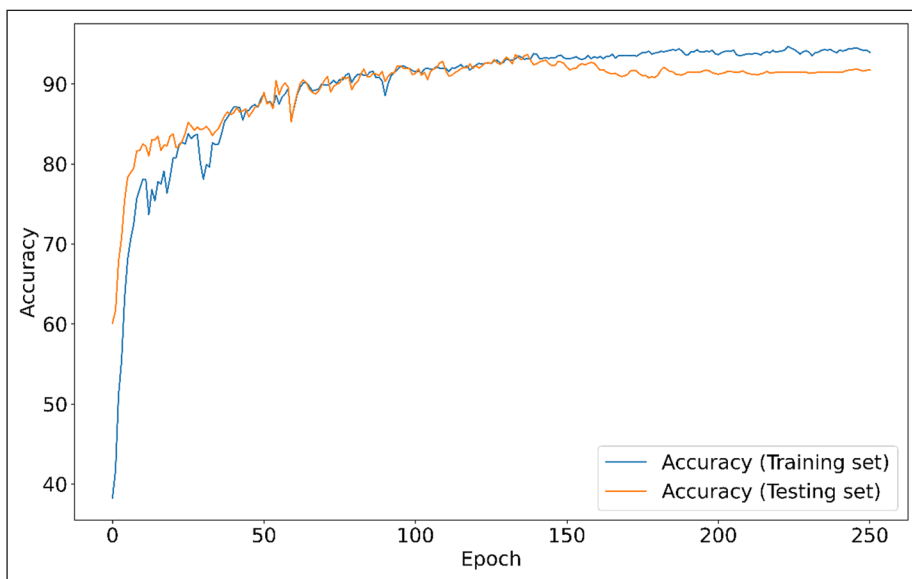


Figure 5 The accuracy of training and testing sets for the group classification dataset.

of two. Finally, accuracy is a general-purpose metric, representing the proportion of correctly classified examples in the overall predictions.

The proposed framework for sub-object classification demonstrated its validity in [Tables 1 and 2](#), with 96% and 93% accuracy, respectively. F1-scores are proximate for all three classes in [Table 1](#), although the sdB class has 25% more training samples, proving the robustness of the model. Moreover, precision and F1-scores improve for all subtypes by eliminating poorly represented classes and addressing model confusion. The results also confirm the method's applicability to the group classification task, where the observed 5% increase in scores can be attributed to architecture refinements, loss function, optimizer, training pipeline, and other advancements in the recent technique. Furthermore, by expanding the sample sizes for corresponding classes, future works can achieve similar accuracies on different groups and sub-groups, such as white dwarfs, M-type stars, Mrk Abs, sdA, and C-N.

Additional experiments have been carried out to explore the DFBS objects further. Two different inference datasets with sizes of 1 million and 4 million objects have been extracted using the unlabeled object extraction three-step algorithm, detailed in [Section 3.1](#). Objects from randomly selected astronomical plates have been extracted to collect datasets of the mentioned sizes while ensuring the diversity of the objects. This way, objects from 65 and 266 astronomical plates were aggregated to assemble the small and large inference datasets. [Tables 3 and 4](#) show the predicted samples for each sub-group under different confidence thresholds. Objects predicted with confidence less than the chosen threshold are assigned to column *Other*. Reportedly, for the selected range of thresholds (0.8–0.95),

13-13.5% of the objects belong to one of the targeted subclasses. Precision scores for each subgroup rise concurrently with threshold values, while recall scores slightly drop. Thereby, classification results converge to actual numbers of sub-objects considering the model's accuracy. Drawing from the prior and generalizing for the whole survey, with the utmost confidence in the model's predictions when the highest threshold is selected, 1/20 of the twenty million astronomical objects can be considered to belong to one of the three studied subgroups. The 1/20 ratio refers to the successful classification of objects into one of the nine subgroups during the training phase of the model. This means that, out of the total 20 million objects, the classification model accurately assigned objects to their respective subgroups for 1/20th of the dataset. Increasing the number of samples for all nine subclasses, presented in Section 1, and performing additional experiments will provide comprehensive insights into the three main groups of the DFBS.

THRESHOLD	C-H	Mrk SB	sdB	OTHER
0.95	10557	10237	15084	964123
0.9	23823	19044	21691	935443
0.85	38633	29163	28265	903940
0.8	56026	40609	35231	868135

THRESHOLD	C-H	Mrk SB	sdB	OTHER
0.95	37355	48695	57278	3856673
0.9	85379	89872	82183	3742567
0.85	139925	134719	106725	3618632
0.8	204382	184918	133209	3477492

Table 3 The model's inference results on one million astronomical objects.

Table 4 The model's inference results on four million astronomical objects.

5 SUMMARY AND CONCLUSION

The paper presents a framework for automated detection and sub-object classification of astronomical objects in the DFBS. The framework combines a four-step image-processing pipeline with deep neural network architectures to achieve promising astronomical sub-object classification task results. The proposed framework is the pioneering work on the image-based sub-object classification of astronomical objects. The framework's effectiveness is demonstrated by evaluating astronomical subtypes with sufficient examples. The proposed framework for sub-object classification demonstrated its validity with more than 96% accuracy. Moreover, it illustrated its superiority on the astronomical group classification task in the DFBS by outperforming the preceding work by 5% accuracy refinement. Additional experiments on 1 million and 4 million inference datasets revealed the object numbers of selected subgroups in the DFBS. According to the findings, 1/20 of the 20 million astronomical objects belong to the subjects of the inner group classification.

Additionally, this work contributes to the astronomical scientific community by providing a general cloud-based service for image-based astronomical object classification. The service offers users a cloud-based infrastructure for different ML experiments on provided and proprietary astronomical data. Training, fine-tuning, evaluation, and inference pipelines are implemented within the service, allowing for integration and execution of the entire ML workflow. Additionally, it enables users to monitor their pipelines using TensorBoard-based visualizations and other performance metrics.

Future work aims to increase the number of sub-object classes by including poorly represented classes into the model and leveraging distributed computing approaches for large-scale inference on the overall First Byurakan Survey database. The authors intend to annotate additional sub-object datasets to address these issues and provide researchers with a more comprehensive toolset for applying to a broader range of astronomical surveys.

FUNDING INFORMATION

The paper is supported by the European Union's Horizon 2020 research infrastructures program under grant agreement No 857645, project NI4OS Europe (National Initiatives for Open Science in Europe), and by the Republic of Armenia Ministry of Education and Science (RA MES) Committee of Higher Education and Science, in the frames of the Advanced Research Projects 21AG-1C053 (2021–2026) and 21AG-1B052.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Hrachya Astsatryan  orcid.org/0000-0001-8872-6620

Institute for Informatics and Automation Problems (IIAP), P. Sevak 1, 0014 Yerevan, Armenia

Stepan Babayan

Institute for Informatics and Automation Problems (IIAP), P. Sevak 1, 0014 Yerevan, Armenia; National Polytechnic University of Armenia (NPUA), 105, Teryan str., 0009 Yerevan, Armenia

Areg Mickaelian

Byurakan Astrophysical Observatory (BAO), 0213 Byurakan, Aragatzotn province, Armenia

Gor Mikayelyan

Byurakan Astrophysical Observatory (BAO), 0213 Byurakan, Aragatzotn province, Armenia

Martin Astsatryan

Rutgers University-New Brunswick, New Brunswick, NJ 08901-8554, USA

REFERENCES

- Abolfathi, B**, et al. 2018. The fourteenth data release of the Sloan digital sky survey: First spectroscopic data from the extended baryon oscillation spectroscopic survey and from the second phase of the Apache Point Observatory Galactic Evolution Experiment. *The Astrophysical Journal Supplement Series*, 235(2): 42. DOI: <https://doi.org/10.3847/1538-4365/aa9e8a>
- Alawi, AEB** and **Al-Roainy, AA**. 2021. Deep residual networks model for star-galaxy classification. In: *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, pp. 1–4. IEEE.
- Arbelaez, P**, et al. 2010. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5): 898–916. DOI: <https://doi.org/10.1109/TPAMI.2010.161>
- Astsatryan, H**, et al. 2021. Astronomical objects classification based on the Digitized First Byurakan Survey low-dispersion spectra. *Astronomy and Computing*, 34: 100442. DOI: <https://doi.org/10.1016/j.ascom.2020.100442>
- Astsatryan, H**, **Shoukourian, Y** and **Sahakyan, V**. 2004. The ArmCluster Project: brief introduction. In: *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications*. PDPTA'04, 1291–1295.
- Bock, S** and **Wei, M**. 2019. A proof of local convergence for the Adam optimizer. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. pp. 1–8. IEEE. DOI: <https://doi.org/10.1109/IJCNN.2019.8852239>
- Burke, CJ**, et al. 2019. Deblending and classifying astronomical sources with Mask R-CNN deep learning. *Monthly Notices of the Royal Astronomical Society*, 490(3): 3952–3965. DOI: <https://doi.org/10.1093/mnras/stz2845>
- Carneiro, T**, et al. 2018. Performance analysis of Google Colaboratory as a tool for accelerating deep learning applications. *IEEE Access*, 6: 61677–61685. DOI: <https://doi.org/10.1109/ACCESS.2018.2874767>
- Cloud-ML**. 2021. *Computational astrophysics library*. Available at <https://github.com/ArmHPC/Computational-Astrophysics>.
- Cui, CZ** and **Zhao, YH**. 2007. Worldwide R&D of virtual observatory. *Proceedings of the International Astronomical Union*, 3(S248): 563–564. DOI: <https://doi.org/10.1017/S1743921308020152>
- Ethiraj, S** and **Bolla, BK**. 2021. Classification of astronomical bodies by efficient layer fine-tuning of deep neural networks. In: *2021 5th Conference on Information and Communication Technology (CICT)*. pp. 1–6. IEEE. DOI: <https://doi.org/10.1109/CICT53865.2020.9672430>
- González, RE**, **Munoz, RP** and **Hernández, CA**. 2018. Galaxy detection and identification using deep learning and data augmentation. *Astronomy and computing*, 25: 103–109. DOI: <https://doi.org/10.1016/j.ascom.2018.09.004>

- Hanisch, RJ**, et al. 2015. The virtual astronomical observatory: re-engineering access to astronomical data. *Astronomy and Computing*, 11: 190–209. DOI: <https://doi.org/10.1016/j.ascom.2015.03.007>
- He, K**, et al. 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778. DOI: <https://doi.org/10.1109/CVPR.2016.90>
- He, K**, et al. 2017. Mask r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2961–2969. DOI: <https://doi.org/10.1109/ICCV.2017.322>
- Heymans, C**, et al. 2012. CFHTLenS: the Canada–France–Hawaii telescope lensing survey. *Monthly Notices of the Royal Astronomical Society*, 427(1): 146–166. DOI: <https://doi.org/10.1111/j.1365-2966.2012.21952.x>
- Huchra, JP**. 1977. The nature of Markarian galaxies. *Astrophysical Journal Supplement Series*, 35: 171–195. DOI: <https://doi.org/10.1086/190474>
- Jia, P, Liu, Q and Sun, Y**. 2020. Detection and classification of astronomical targets with deep neural networks in wide-field small aperture telescopes. *The Astronomical Journal*, 159(5): 212. DOI: <https://doi.org/10.3847/1538-3881/ab800a>
- Kim, EJ and Brunner, RJ**. 2016. Star-galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, stw2672. DOI: <https://doi.org/10.1093/mnras/stw2672>
- Kingma, DP and Ba, J**. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Mickaelian, AM**, et al. 2007. The digitized first Byurakan survey–DFBS. *Astronomy & Astrophysics*, 464(3): 1177–1180. DOI: <https://doi.org/10.1051/0004-6361:20066241>
- Mickaelian, AM**, et al. 2016. Ten years of the Armenian Virtual Observatory. *Astronomical Surveys and Big Data*, 505: 16.
- Mickaelian, AM**, et al. 2023. Armenian virtual observatory: status and activities. *Astronomy and Computing*, 42: 100689. DOI: <https://doi.org/10.1016/j.ascom.2023.100689>
- Pan, SJ and Yang, Q**. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359. DOI: <https://doi.org/10.1109/TKDE.2009.191>
- Pavlidis, T**. 1980. Algorithms for shape analysis of contours and waveforms. *IEEE Transactions on pattern analysis and machine intelligence*, 4: 301–312. DOI: <https://doi.org/10.1109/TPAMI.1980.4767029>
- Quinn, PJ**, et al. 2004. The International Virtual Observatory Alliance: recent technical developments and the road ahead. In: Quinn, PJ and Bridger, A (eds.), *Optimizing Scientific Return for Astronomy Through Information Technologies 5493*. Glasgow: SPIE. pp. 137–145. DOI: <https://doi.org/10.1117/12.551247>
- Sandler, M**, et al. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4510–4520. DOI: <https://doi.org/10.1109/CVPR.2018.00474>
- Tran, GS**, et al. 2019. Improving accuracy of lung nodule classification using deep learning with focal loss. *Journal of healthcare engineering*, 2019: 57–65. DOI: <https://doi.org/10.1155/2019/5156416>
- York, DG**, et al. 2000. The Sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3): 1579. DOI: <https://doi.org/10.1086/301513>

TO CITE THIS ARTICLE:

Astsatryan, H, Babayan, S, Mickaelian, A, Mikayelyan, G and Astsatryan, M. 2024. Cloud-Based Machine Learning Service for Astronomical Sub-Object Classification: Case Study On the First Byurakan Survey Spectra. *Data Science Journal*, 23: 6, pp. 1–14. DOI: <https://doi.org/10.5334/dsj-2024-006>

Submitted: 04 October 2023

Accepted: 04 January 2024

Published: 30 January 2024

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.