IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 14, 2020, accepted March 2, 2020, date of publication March 5, 2020, date of current version March 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978804

Galaxy Image Classification Based on Citizen Science Data: A Comparative Study

MANUEL JIMÉNEZ[1], MERCEDES TORRES TORRES[2], ROBERT JOHN[1], (Senior Member, IEEE), AND ISAAC TRIGUERO[1], (Member, IEEE)

[1]Computational Optimisation and Learning Laboratory (COL), School of Computer Science, University of Nottingham, Nottingham NG8 1BB, U.K.
[2]Computer Vision Laboratory (CVL), School of Computer Science, University of Nottingham, Nottingham NG8 1BB, U.K.

Corresponding author: Manuel Jiménez (manuel.jimenezmorales@nottingham.ac.uk)

This work was supported in part by the NVIDIA Corporation, and in part by the Titan Xp GPU. The work of Manuel Jiménez was supported by the School of Computer Science, University of Nottingham, through the Ph.D. Scholarship.

ABSTRACT Many research fields are now faced with huge volumes of data automatically generated by specialised equipment. Astronomy is a discipline that deals with large collections of images difficult to handle by experts alone. As a consequence, astronomers have been relying on the power of the crowds, as a form of citizen science, for the classification of galaxy images by amateur people. However, the new generation of telescopes that will produce images at a higher rate highlights the limitations of this approach, and the use of machine learning methods for automatic classification is considered essential. The goal of this paper is to shed light on the automated classification of galaxy images exploring two distinct machine learning strategies. First, following the classical approach consisting of feature extraction together with a classifier, we compare the state-of-the-art feature extractor for this problem, the WND-CHARM, with our proposal based on autoencoders for feature extraction on galaxy images. We then compare these results with an end-to-end classification using convolutional neural networks. To better leverage the available citizen science data, we also investigate a pre-training scheme that exploits both amateur- and expert-labelled data. Our experiments reveal that autoencoders greatly speed up feature extraction in comparison with WND-CHARM and both classification strategies, either using convolutional neural networks or feature extraction, reach comparable accuracy. The use of pre-training in convolutional neural networks, however, has allowed us to provide even better results.

INDEX TERMS
Astroinformatics, autoencoders, citizen science, convolutional neural networks, deep learning, feature extraction, galaxy morphologies, image classification.

I. INTRODUCTION

Classification is one of the core tasks addressed by machine learning (ML) algorithms [1], [2]. A classifier is usually trained to learn patterns from input data, aiming to predict the label to be assigned to previously unseen data instances [3]. In image classification, we pursue the categorisation of images into two or more classes, being either mutually exclusive (multi-class classification, in which only one single class is assigned) or not (multi-label classification, where different classes coexist). These paradigms are widely implemented in multiple real-world applications such as fingerprint identification [4] or the recognition of facial emotions [5], and they have also become a useful tool for data analysis in science and engineering [6], [7].

Astronomers have seen their data processing capabilities exceeded with the advent of modern instrumentation [8], leading to the emergence of the *astroinformatics* discipline [9] to help analyse the data provided. In most cases, this entails the classification of large collections of astronomical images [10]–[12]. Particularly, the morphological classification of galaxy images aims at the categorisation of these objects into two main classes (morphologies), elliptical and spiral [13]. The morphology is a key indicator for understanding the galaxy inner structure and physical processes, also revealing aspects about the formation and evolution of the universe [14]. However, due to the huge amounts of images produced in modern telescopes [15], astronomers

The associate editor coordinating the review of this manuscript and approving it for publication was Shun-Feng Su.

have been drawing upon the general public for this task using the internet, giving rise to the re-emergence of the citizen science movement [16], [17]. This was first materialised with the release of the Galaxy Zoo 1 (GZ1) project [18], which generated the largest manually annotated catalogue of galaxy images to date [19]. Nonetheless, the next generation of astronomical surveys that will produce billions of galaxy images [20] shows the limitations of this approach. ML methods are needed, pursuing a robust automation of the classification task, and several efforts have recently been developed in this direction [21]–[24].

The traditional ML approach to image classification requires the extraction of features from the image. Classical learning algorithms (e.g. Decision Trees, k-Nearest Neighbours) do not cope well with images directly, which is typically solved transforming the image pixels into a new feature space by means of feature extraction (FE) techniques [25]. In contrast, deep learning (DL) based approaches [26] need minor or no data preprocessing for the classification of images. More specifically, convolutional neural networks (CNNs) [27], [28] provide excellent solutions [29], being capable of taking a raw image as input and perform an implicit FE process along with the classification in one single step. Recent state-of-the-art CNNs are usually composed of a very large number of layers [30] when dealing with challenging image classification problems [31]. Alternatively, DL can also be used to extract features of an image by means of autoencoders (AEs) [32], which have also been proposed to ease the learning of standard classifiers [33]. Whereas CNNs often need to learn the image features from scratch using a large amount of labelled data, AEs enable the encapsulation of the FE process for a particular problem without any need of labels, which can be advantageous for the classification of big collections of images and the use of other kind of classifiers [33].

The classification of galaxy images has leveraged both general strategies [34]. However, the variable characteristics of the images commonly used in the training of ML for this problem have systematically neglected a fair comparison of different methods under the same learning conditions. This work presents a comparative study of distinct approaches for galaxy image classification, investigating their advantages and disadvantages in a common experimental framework. Following the classical approach, we explore the suitability of two feature extractors. On the one hand, we take the WND-CHARM multi-purpose feature extractor [35] as the state-of-the-art FE method [22], [36], [37] used in this problem. On the other hand, we propose two AE architectures for the FE of this kind of images, which, to the best of our knowledge, has not been explored yet. We also analyse the effect of two feature selection methods on the resulting feature sets, and then compare these results with an end-to-end approach using CNNs. Here we propose a simple yet effective CNN architecture and compare it with a deeper CNN, namely ResNet [31].

The experiments are carried out over the GZ1 main dataset, consisting of nearly 668k images annotated by amateurs for which we also hold expert classifications for a subset of ∼41k examples. We first explore the influence in the results of several factors such as the image size or the presence of colour channels using the subset with expert classifications. We then investigate the scalability of both classification approaches in the larger dataset. Finally, we utilise the whole GZ1 data with both label sets to pre-train the CNNs using amateur classifications and then fine-tune them on the expert-labelled subset.

The rest of the paper is structured as follows. In Section II, we briefly introduce related work about the classification of galaxy images with citizen science and the ML approaches taken for this problem to date. Section III presents our proposed models of AEs and CNN for the FE and classification of galaxy images, respectively, including a brief explanation of the proposed pre-training approach with citizen science data. Then, in Section IV we explain the experimental setup established. Section V presents the results and discussion, and finally Section VI concludes the paper.

II. RELATED WORK

This section provides information about the central concepts of the paper. First, we describe in more detail the classification of galaxy images with citizen science (Section II-A). After this, we present related work about ML approaches for the automatic classification of galaxy images (Section II-B).

A. GALAXY IMAGE CLASSIFICATION AND CITIZEN SCIENCE

Here we consider the morphological classification of galaxy images, that is, the classification of these objects according to a blend of the galaxy shape, colour, and texture [38]. This has been standard practice since it was first applied by E. Hubble nearly a century ago [13]. The morphology provides a first-order descriptor about the galaxy, which is key for astronomers in the study of fundamental questions about their inner physics [39], interactions [40], or evolution [41]. There are two main morphological types based on the presence or not of a disk: spiral and elliptical, respectively. However, the multiplicity of hybrid types and the wide range of image conditions depending on factors such as the galaxy brightness, size or distance, turn the classification of this sort of images into a very complex task.

Citizen science has been a partial solution for this problem, with the engagement of amateur people from the general public in this kind of data analysis [42]. Citizen science projects join the endeavours of myriads of volunteers committed to helping with a task that typically is time consuming and tedious for experts, but also decisive for getting advances in a certain research problem. A task usually covered is the classification of images, for which the Galaxy Zoo project represents the most successful implementation to date [43]. Its first edition, the Galaxy Zoo 1 (GZ1) [18], was focused on the distinction among spiral and elliptical morphologies, providing amateur classifications for nearly 900k galaxy images. In addition, there are also expert classifications for a

subset of ∼41k of the GZ1 images, as it is explained in [18]. This inclusion of amateurs in research tasks has brought additional uncertainty into results. Nonetheless, given the potential of this large-scale data processing from the ML perspective, solutions are being proposed to overcome this issue [44].

Along with numerous scientific insights,[1] GZ1 and the subsequent releases of the project have also generated enough labelled data for the proper training of ML algorithms [19], [45], [46]. However, ML implementations using this data have generally aimed at replicating participants' classification skills [21], [23] instead of tackling the uncertainty in the results and/or taking advantage of available expert classifications [18]. In this work, we are interested in exploring how the inclusion of amateur labels affects the classification results, leveraging this particular characteristic featured by the GZ1 dataset.

B. MACHINE LEARNING STRATEGIES FOR GALAXY IMAGE CLASSIFICATION

This section presents an overview of works found in the specialised literature concerning the use of ML for galaxy image classification. First, FE based approaches are reviewed (Section II-B.1). Then, we briefly introduce CNNs and their latest trends, and we examine their use in galaxy image classification (Section II-B.2).

1) FEATURE EXTRACTION PLUS A CLASSIFIER

FE methods used in the classification of galaxy images can be grouped into two main categories: problem specific, which have been especially devised for this particular problem, and general, which cope with image classification regardless of the problem definition. Among the first, we review the use of physical parameters extracted from the image, whereas the second category is dominated by the WND-CHARM feature set.

The classification of galaxy images with ML started with the extraction of a reduced number of physical parameters from the image [47], [48]. These parameters accounted for properties such as galaxy ellipticity, surface brightness, or concentration. As a form of image features, they were then classified using artificial neural networks [49] (ANNs). The feature set was then extended to a greater number of parameters and standardised with the so-called CAS (Concentration-Asymmetry-Smoothness) methods [50], [51]. ANNs were generally used for the classification, along with other general classifiers such as random forests and support vector machines [52], [53]. Nonetheless, with the improvement of computational resources, general purpose feature extractors that compute longer sets of features have outperformed these first attempts.

Although it was originally developed as an image analysis tool for the classification of biological images [35], WND-CHARM represents the state-of-the-art feature extractor for the classification of galaxy images [22], [36], [37]. First, through a FE phase, it computes a set of families of features from the raw images in a one by one fashion. These are categorised as image content descriptors, image transforms, and compound image transforms (transformations of a previous image transformation), composing a feature vector for the image at hand [54]. Depending on the presence or not of colour in the image, two feature sets are available. Then, a feature selection (FS) phase chooses the most informative subset by calculating the Fisher discriminant score of each feature. Finally, the resulting feature vectors are classified using a modified nearest neighbour (NN) rule [55] that weights the distances using the Fisher scores previously computed. Whereas in traditional NN only the closest (or k closest) examples determine the class, with WND-CHARM the distances to all training samples of each class are measured. In this work, we only consider the FE and FS phases of the WND-CHARM, the so-called WND-CHARM feature map, for the comparison of this method with the FE performed with AEs.

2) CONVOLUTIONAL NEURAL NETWORKS

CNNs have systematically outperformed the classical benchmarks in image classification in the last few years [31], [56], thanks to the ease of access to big datasets and the improvements in computational resources. This has also been the case in galaxy image classification, promoted by the wide availability of astronomical surveys on the web [15]. Basically, CNNs are ANNs with many hidden layers that progressively reach more abstract representations of the input data by computing non-linear transformations. These layers are generally one of these types: convolution layer, pooling layer, and fully connected layer. Whereas convolution and pooling layers build and shrink the feature maps, respectively, fully connected layers try to learn the global information present at the end of these processes [28].

In the recent literature, very deep networks with a large number of layers have been investigated [56]. The network depth is of crucial importance for challenging image classification problems [30]. Many deep neural network architectures (which use hundreds of layers) such as ResNet [31], ResNext [57] or HRnet [58], have provided an outstanding performance in varied image datasets with multitude of different objects [59]. These complex architectures are usually exploited in a pre-trained fashion [60], which saves computational efforts and allows different domains to take advantage of their prediction capabilities when the scarcity of annotated examples invalidates the training of such models from scratch [61]–[63].

CNNs have also been widely used in the classification of galaxy images, showing the limitations of FE methods when the classification is not restricted to the two main morphologies. One of the first successful implementations took place in the framework of a Kaggle competition,

[1]The complete list of peer-reviewed publications based on the Galaxy Zoo project results is available at https://www.zooniverse.org/about/publications.

the Galaxy Challenge,[2] which aimed at classifying a sample of ∼50,000 galaxies from the Galaxy Zoo 2 dataset [45]. The goal was to predict the participants answers to a set of questions about morphological traits featured by the galaxy. The winner CNN architecture [23] established a benchmark for this problem that has been widely employed thereafter [24], [64]. However, these models make use of datasets of moderate size to make feasible their training and employ larger resources in terms of computational means and runtime.

In this work, we propose a simpler CNN architecture to distinguish between the two main morphological types. We test this model against a well-established deeper model, ResNet [31], and explore how their performances are affected by the number and size of the images as well as the presence or not of colour channels. We then compare these results with the classification using FE plus classifier, aiming to investigate in which occasions the different approaches work better. Additionally, we also explore the pre-training of both CNN models considered by exploiting the availability of expert and amateur labels within citizen science data, which to the best of our knowledge has not been investigated before.

III. METHODOLOGY: GALAXY IMAGE CLASSIFICATION WITH DEEP LEARNING

This section introduces the proposed AE architectures for developing the FE of galaxy images, as well as the CNN proposed for comparison of ML approaches for galaxy image classification. First, we give a brief introduction to AEs and describe the two models used through the experiments (Section III-A). Then, we present the used CNN architecture (Section III-B). Finally, we provide some background about the pre-training of CNNs and explain our novel approach to make use of amateur and expert labels (Section III-C).

A. AUTOENCODERS FOR FEATURE EXTRACTION ON GALAXY IMAGES

AEs are a common architecture in unsupervised DL, the referred as *encoder-decoder* [65]. Basically, an AE is an ANN able to build new encodings for some input by means of a symmetrical structure of layers that tries to resemble the input pattern to the output as closely as possible [32]. The middle layer (symmetry axis) represents the encoding, which is found after a training process that does not make any use of the data labelling. As with other ANNs, the set of weights and activation function associated with every neuron generate the outputs layer by layer. A loss function computes the disagreement between input and reconstruction, which is optimised using the stochastic gradient descent [66] in conjunction with the back-propagation algorithm [67].

AEs have been proposed for diverse tasks related to feature fusion [32] and dimensionality reduction to facilitate the learning of canonical classifiers [33]. In astronomy, AEs show a great potential for the processing and storage of large datasets of images. First, their training is unsupervised, which is key due to the scarcity of reliable image labels. Additionaly, they enable the encapsulation of the FE phase for the exploration of patterns in the data and large-scale storage and management of astronomical images. However, to the best of our knowledge, a comparative study of their use for FE on galaxy images has not been accomplished.

[2]https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

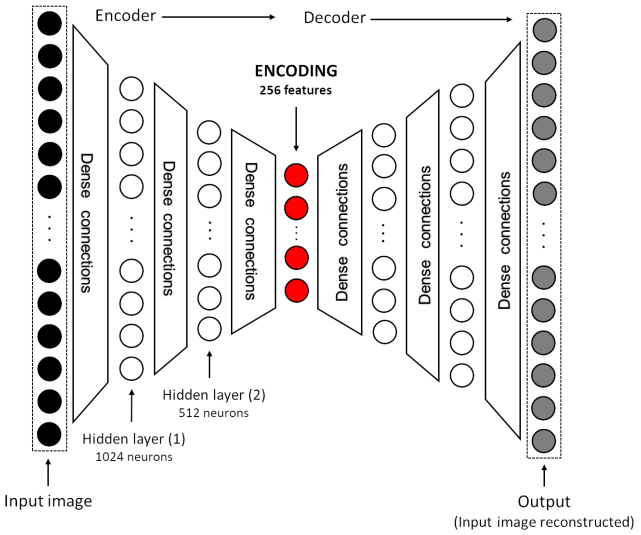


FIGURE 1. Architecture of the deep autoencoder (DAE) proposed.

TABLE 1. Topology of the DAE proposed.

Layer type	Size (neurons)	Activation
Fully connected	1024	ReLU
Fully connected	512	ReLU
Fully connected	**256** (encoding)	ReLU
Fully connected	512	ReLU
Fully connected	1024	Sigmoid

In our comparative study, we implement two different AE models that showed the best performance among a wide set of topologies that were tested. The first one uses fully connected layers in a more classical approach, while the second implements a CNN-based architecture. These models are based on architectures originally designed for the classification of the MNIST dataset [68].

- The Deep AE (DAE) model deploys the architecture of a deep and undercomplete AE, that is, an AE with more than one hidden layer and the encoding having a lower dimensionality than the input [32]. It holds two fully connected layers between input and encoding. Following this, the reverse structure is deployed from encoding to output layer, thus completing the symmetrical structure as it is shown in Figure 1. For this model, the encoding dimension is the same regardless of the input image size: 256 features. The activation function used in all neurons is the rectified linear unit (ReLU) [69], except for the last layer (output), which applies the sigmoid function. The DAE architecture is specified in Table 1.
- The Convolutional AE (CAE) model uses the mechanism of CNNs for learning the encoding, implementing

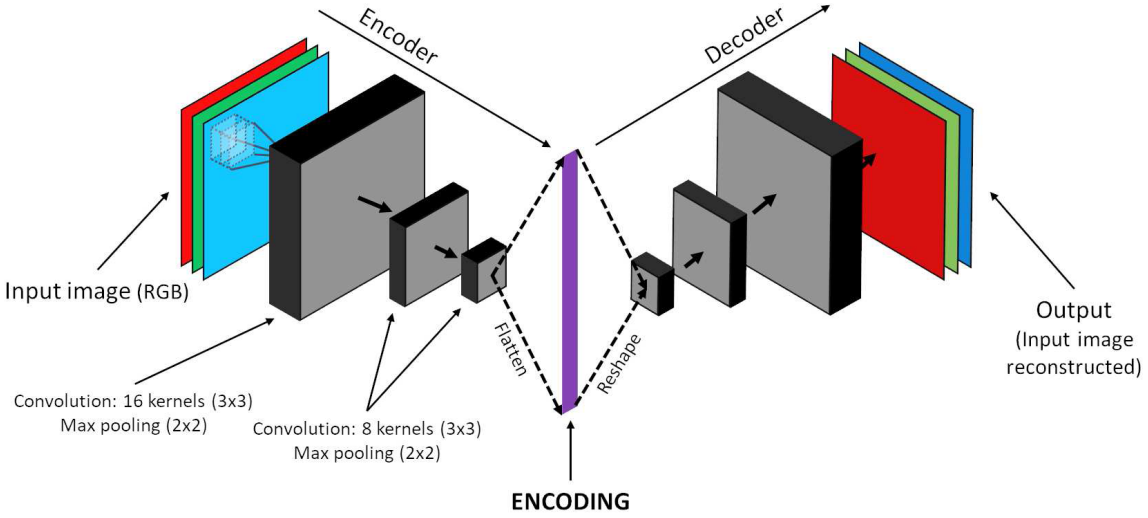


FIGURE 2. Architecture of the convolutional autoencoder (CAE) proposed.

TABLE 2. Topology of the CAE proposed. The reverse structure is deployed for the decoder, using the sigmoid function in the last layer.

Layer type	Size	Stride	Activation
Convolution	16 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–
Convolution	8 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–
Convolution	8 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–

convolution and pooling layers. The model proposed deploys three pairs of convolution – pooling layers from input to encoding. After the third pooling layer, the resulting tensor is flatten to obtain the encoding, as Figure 2 shows. Unlike the DAE model, for CAE the number of features depends on the input image size. The first convolution layer holds 16 kernels and the remaining two hold 8 kernels. The receptive fields are 3×3 pixels size, and pooling layers implement max pooling with 2×2 pixels windows. As in the DAE, ReLU activation functions are used along the network except in the output layer, which applies the sigmoid function. The complete specifications are presented in Table 2.

B. CONVOLUTIONAL NEURAL NETWORK FOR GALAXY IMAGE CLASSIFICATION

Pursuing a fair comparison with the classification using the two AE models introduced above, the CNN architecture proposed here resembles the CAE topology presented in the previous section. Hence, this consists of three consecutive pairs of convolution – pooling layers, which performs the FE phase, followed by three fully connected layers that complete the features classification. As with the CAE, the network computes 16 feature maps in the first convolution layer, and then 8 in second and third convolution layers. Pooling layers implement the max pooling using 2×2 pixels windows. After this, two layers of 256 and 128 neurons hold dense

TABLE 3. Topology of the proposed CNN.

Layer type	Size	Stride	Activation
Convolution	16 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–
Convolution	8 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–
Convolution	8 kernels (3×3)	1 (Zero padding)	ReLU
Pooling	(2×2)	2	–
Fully connected	256	–	ReLU
Fully connected	128	–	ReLU
Fully connected	2	–	Softmax

connections and produce the network output, consisting of two neurons (binary classification). These output layer gives the class probabilities, which are rounded to produce the final classes labels. ReLU activation functions are used through the whole structure except for the output layer, which applies the softmax function that enables us to obtain probability distributions. The architecture proposed is shown in Figure 3, and all specifications presented in Table 3.

Deeper architectures were tested prior to the experiments, implementing up to six pairs of convolution – pooling layers and/or different numbers of feature maps. Nonetheless, the improvement was marginal or even diminished the accuracy in the results. Consequently, we opted to select the CAE's topology presented above and compare this simpler architecture against one of the state-of-the-art models more widely used in computer vision. Particularly, we selected the lighter implementation of ResNet, *ResNet50* [70], which is composed of fifty layers to exploit the residual learning blocks that characterise this DL approach [31].

C. PRE-TRAINING WITH CITIZEN SCIENCE DATA

The final stage of our comparative study involves the pre-training and fine-tuning of the proposed CNN and the ResNet using citizen science data. By this, we aim to investigate the learning of such models from both amateur and expert

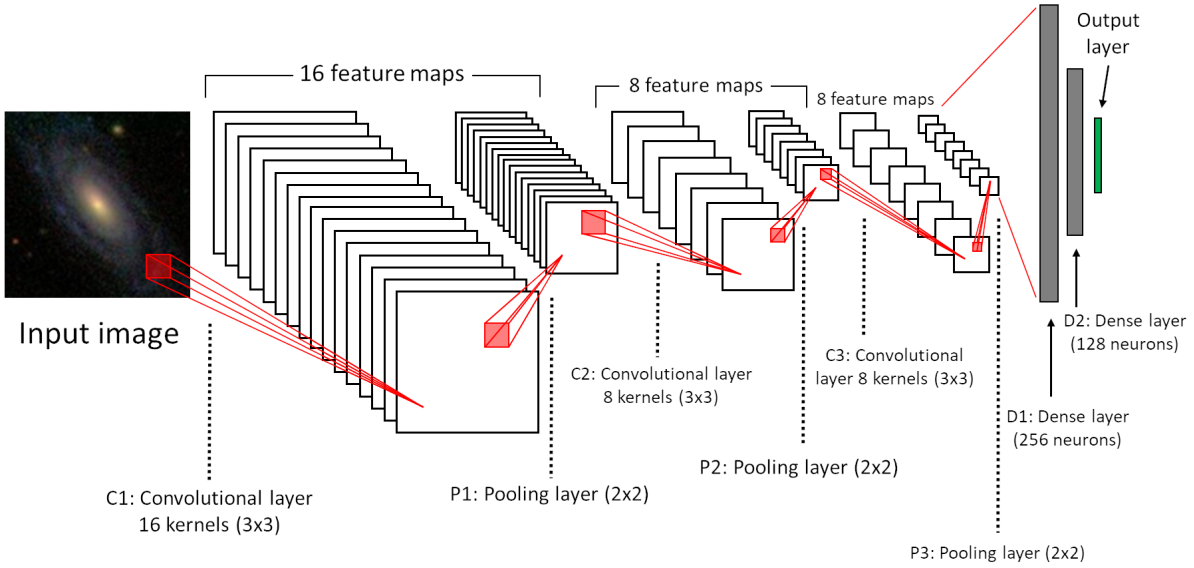


FIGURE 3. Architecture of the proposed CNN.

labels at the same time, thus leveraging the whole range of knowledge featured by citizen science projects.

In the literature, pre-training has been used to alleviate the scarcity of labelled examples and also save the higher computational capacity and time required for the training of deep architectures [63]. Hence, pre-trained networks, which have been previously *fully-trained* using a more generic dataset, are adapted (*fine-tuned*) to the particular classification problem at hand employing more specific data [61]. In practical terms, this process generally entails the top dense layers of the network that performs the classification of the deep features extracted by the convolutional layers. Taking advantage of the set of weights previously learned in the pre-training step, the dense layers are modified and trained to fulfil the requirements of the targeted classification problem [70].

Inspired by the problem of coarse supervision [71], in which image labels followed a hierarchical structure (e.g. coarse: big cat, finer: leopard, cheetah, tiger), here we explore a special case of pre-training by taking advantage of citizen science data. First, we train the CNN using amateur-labelled data, which is more abundant with respect to expert-labelled data and considered less reliable [44]. Then, we complete the fine-tuning with examples labelled by experts, which we take as ground truth for this problem.

IV. EXPERIMENTAL SETUP

In this section, we explain the experimental setup deployed for the comparative analysis of the classification approaches introduced above. First, we present the selected standard classification algorithms (Section IV-A) and the two chosen FS methods (Section IV-B), along with their parameters. Both standard classifiers and FS methods are applied to either AEs or WND-CHARM feature sets. Then, we provide the experimental details of the DL-based models, which include both AEs, the proposed CNN, and ResNet (Section IV-C). Finally, we introduce the GZ1 image data and the performance evaluation of the experiments (Section IV-D).

A. STANDARD CLASSIFICATION ALGORITHMS

The standard classification algorithms selected for the comparative study were made according to their proven accurate behaviour in many real-world applications [72]. Here we use k-nearest neighbours, random forests, and support vector machines. The basics of each algorithm and their configurations are summarised in Table 4. We are not interested in performing a fine tuning of these algorithms, and we thus employ the standard parameters.

- The k-nearest neighbours (kNN) [55] is a well-known algorithm in supervised learning. The kNN algorithm uses the whole training set as to classify new instances via a similarity function, which is usually defined by a distance on the feature space. First, distances are computed from the new example to the entire training set, and then the k closest training instances are selected. The final label is chosen in accordance to the prevalent class in the k-subset. Here, we use the Euclidean distance and $k = 3$.
- Random forests (RF) [73] is a classifier based on the decision tree algorithm [2]. It trains a number n of decision tree classifiers and provides the majority class among them as a result. Each sub-tree, referred as an estimator, is trained using a sub-sample of the original training set. Here we run the experiments with $n = 100$. To measure the quality of a split, the Gini impurity is used.
- Support vector machine (SVM) [74] is capable of learning a mapping from the input attributes to the set of classes by means of a higher-dimensional feature space. For this, a kernel function enables the computations of the inner product between two feature vectors. After the

TABLE 4. Parameters for the selected standard classifiers.

Algorithm	Parameters
kNN	$k = 3$ Distance: *Euclidean*
RF	$n = 100$ Criterion: *gini*
SVM	Kernel type: *Linear* $C = 1.0$, Tolerance parameter = 10^{-4}

instances are mapped to the new feature space, the algorithm looks for the optimal separating hyperplane, that is, the one that maximises the distance to the different classes clusters. Here we train the SVM using the linear kernel.

B. FEATURE SELECTION METHODS

We consider two feature selection (FS) methods through the comparative study, aiming to investigate the impact of FS over the extracted feature sets in terms of classification runtime and accuracy:

- As mentioned in Section II-B.1, a FS method is proposed as part of the WND-CHARM classifier [35] and is based on a feature ranking involving the use of Fisher discriminant scores as feature weights [75]. First, the weights are calculated for each feature following Equation (1):

$$W_f = \frac{\sum_{c=1}^{N}(\overline{T_f} - \overline{T_{f,c}})^2}{\sum_{c=1}^{N}\sigma_{f,c}^2} \cdot \frac{N}{N-1}, \tag{1}$$

  where W_f is the Fisher score of feature f, N is the number of classes in the problem, $\overline{T_f}$ is the mean of the values of feature f in the entire dataset, $\overline{T_{f,c}}$ is the mean of the values of feature f in the class c, and $\sigma_{f,c}^2$ is the variance of feature f values across all examples of class c. After the features are ranked according to this weighting, the 35% holding a lower Fisher score are rejected, as originally proposed in [35]. Therefore, this Fisher FS method results in a fixed number of features.
- The second FS method proposed is part of the so-called embedded FS techniques, which highlights as a simple yet fast strategy especially suitable for high-dimensional data [76]. It implements a randomised decision tree with 50 trees to choose the most relevant features. This method also makes use of the labels, searching an optimal subset of features in the combined space of features and hypotheses. As such, the final number of features selected is variable, depending on the data sample considered [77].

C. DL-BASED MODELS

In this paper we pursue a fair comparison between distinct approaches for the classification of galaxy images. Thus, as stated above, the DL-based models used in the experiments share a similar architecture and most parameters. Nonetheless, to compare the proposed CNN against a well-established deep neural network for end-to-end image classification, we introduce the ResNet [31] model as a comparison algorithm. ResNet is an architecture normally trained to distinguish between multiple classes. Given that we focus on a binary classification problem, we use one of the lightest versions of this network, the *ResNet50* [70].

AE models, the proposed CNN and ResNet were trained over 100 epochs with a batch size of 256 examples. CAE, CNN and ResNet were optimised with stochastic gradient descent [77], whereas DAE used the adadelta optimiser [78].

D. DATA AND EVALUATION OF EXPERIMENTS

The data used in the experiments is part of the collection of galaxy images classified by the GZ1 project, which results were published in the so-called GZ1 Table 2 (GZ1-T2) after the project closure[3] [19]. This dataset includes amateur classifications for a total of 667,944 galaxy images. In addition, we also hold expert classifications for a subset of the GZ1-T2 data. This sample, referred from now on as GZ1 Expert subset (GZ1-E), comprises an amount of 41,424 examples that were classified as elliptical or spiral by a team of expert astronomers [18].

The entire GZ1-T2 image dataset was primarily downloaded from the Sloan Digital Sky Server (SDSS) CAS server.[4] In order to establish a fair comparison, we follow the original GZ1 project specifications [18], taking 423 × 423 pixels JPEG images centred in the galaxy. The image scale is particular for each image and varies in accordance to the formula $0.024R_p$ arcsec/pixel, where R_p is the Petrosian radius for the galaxy, that is, a good estimator of its physical size. However, we found that this automatic scaling tends to leave the galaxy isolated in the centre of the image, with a dominance of background pixels and/or other meaningless artifacts around the target object. Therefore, in order to speed up the image processing by both FE methods and the CNNs, we simplified the images' presentation and ended up considering two image sizes through the experiments, aiming to study the influence of the image size and resolution in the classification performance: 128 × 128 (128x) and 64 × 64 (64x) pixels images. To accomplish this, we first cropped the original images in the GZ1-T2 dataset to their half size (212 × 212 pixels) and converted them to TIFF format, keeping the galaxy in the centre of the image. After this, we compressed the resulting images to the two sizes referred above.

For the GZ1-E, we used the expert classifications available as image labels, which we take as ground truth for the problem. In contrast, the GZ1-T2 data provides the record of votes for the options offered on the GZ1 web to project participants [18]. Hence, when using the whole data we assigned the majority voted class among spiral and elliptical, considering the original amateur classifications. However, this criterion left 8,759 examples for which both scores coincided and thus could not be labelled in this way. We opted to remove these

[3]The GZ1 results are available at http://data.galaxyzoo.org.
[4]http://cas.sdss.org

TABLE 5. Description of the two data samples taken from the GZ1-T2 dataset used in the experiments.

Class	**GZ1 Expert** (GZ1-E)	**GZ1 Amateur** (GZ1-A)
Elliptical	16,375 (39.5%)	430,562 (65.32%)
Spiral	25,049 (60.5%)	228,623 (34.68%)
Total	41,424	659,185

images from the GZ1-T2 data for consistency, then using the remaining 659,185 images. We refer to this sample as GZ1 Amateur subset (GZ1-A) from now on. The classes distribution of both GZ1-T2 data samples are shown in Table 5.

In all the experiments, we used a 5-fold cross-validation scheme. For AEs, CNN and ResNet, the training set was split into 70/30 for training and validation. In the classification with extracted feature sets, the classifiers training were carried out consistently: same data partitions defined for the AEs training were used for the training and testing of the classifiers, aiming to resemble real working conditions for the classification of unseen data.

Since the problem classes are balanced (Table 5), we drew upon the classification rate or accuracy (Acc) measure, which accounts for the proportion of correct classifications with respect to all classified examples [2]. We took as final measure the average over the five test data partitions. We also analysed the performance in terms of runtime, aiming to estimate a runtime comparison between the different approaches studied here. We examined the FE runtime, taken by both AE models and WND-CHARM, and the classification runtime employed by the classifiers, including the CNN and ResNet. For AEs, the runtime shown accounted for the training and computation of features, whereas WND-CHARM directly computed the features with no training. The classification runtime presented accounts for the training and classification stages, considering negligible the time taken by the FS phase, when applied.

All experiments involving the classification of the feature sets were carried out in a single node with an Intel(R) Xeon(R) CPU E5-1650 v4 processor (12 cores) at 3.60GHz, and 64 GB of RAM. For the training of the DL-based models, we employed a NVIDIA Titan Xp GPU. In terms of software, the Keras[5] Python package was used for the AEs, CNN and ResNet, and the Scikit-learn[6] library for all experiments involving either the training and classification phases of the standard classifiers introduced above. The WND-CHARM implementation used here is freely available in Python language at https://github.com/wnd-charm/wnd-charm.

V. RESULTS AND ANALYSIS

This section presents the experimental results. First, we carry out a comparative study using the GZ1-E subset and considering both image sizes as well as colour and greyscale images (Subsection V-A). Then, we focus on the GZ1-A subset to investigate the scalability of the best classifier from GZ1-E, and to leverage the larger number of amateur-based classifications with our proposed pre-training scheme for CNNs (Subsection V-B). Finally, we analyse the results obtained (Section V-C).

[5]https://keras.io/

[6]https://scikit-learn.org

TABLE 6. Number of extracted features in GZ1-E.

	No. Features	
FE method	64x images	128x images
DAE	256	256
CAE Greyscale	512	2,048
CAE Colour	512	2,048
WND-CHARM Greyscale	2,919	2,919
WND-CHARM Colour	4,059	4,059

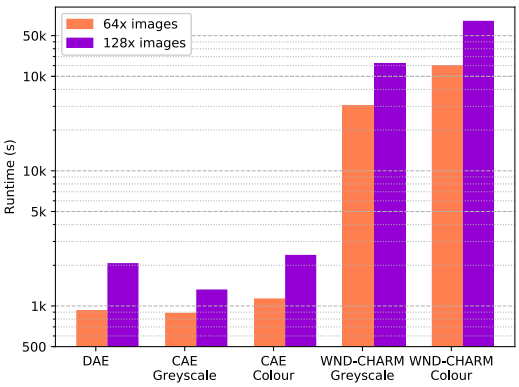


FIGURE 4. FE runtime in logarithmic scale for GZ1-E sample.

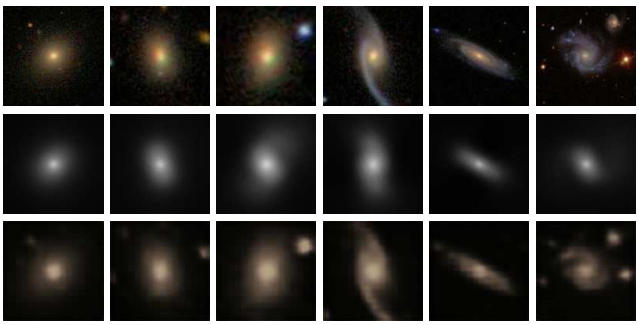


FIGURE 5. Sample of galaxy images reconstructed by the proposed AE models. Top row presents the original 128x images from the GZ1-T2 dataset. Middle and bottom rows show their reconstructions performed by DAE and CAE architectures, respectively.

A. GZ1-E: EXPERT SUBSET

Due to its reduced size, we first used the GZ1-E sample with expert labels, investigating the performances of both AE models proposed as well as the influence of image size and use of colour in classification results. The CAE was tested with colour and greyscale images, whereas the DAE was only tested with greyscale images. For WND-CHARM, the two feature sets available for colour and greyscale images were computed. Thus, we obtained a distinct feature set for each FE method and image colour/size configuration.

The resultant number of features is indicated in Table 6, and a visual comparison of the runtime spent by these methods is presented in Figure 4. For visual illustration of the behaviour of both AEs, Figure 5 plots the image

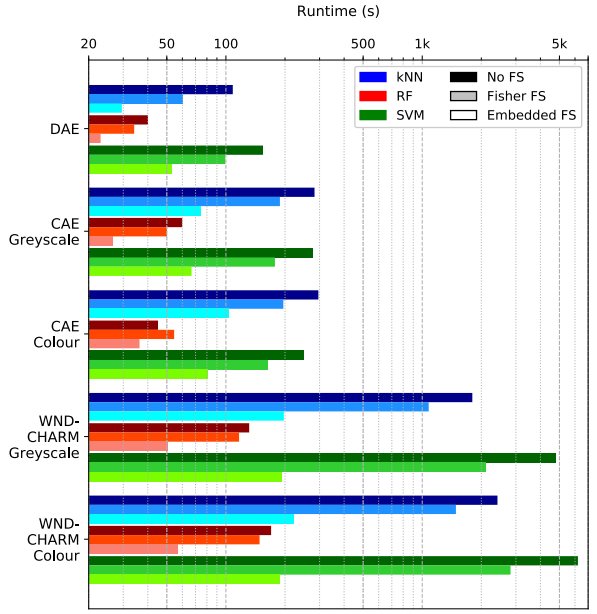


FIGURE 6. Classification runtimes in logarithmic scale for 64x images of the GZ1-E sample. Classifiers and FS methods are represented by colour and intensity, respectively.

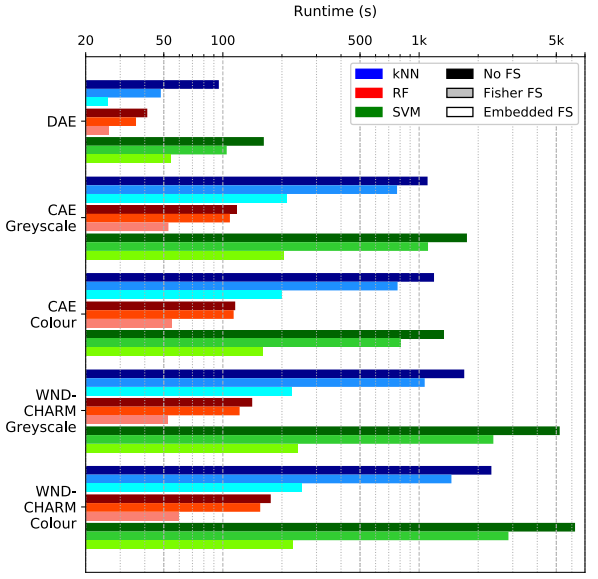


FIGURE 7. Classification runtimes in logarithmic scale for 128x images of the GZ1-E sample. Classifiers and FS methods are represented by colour and intensity, respectively.

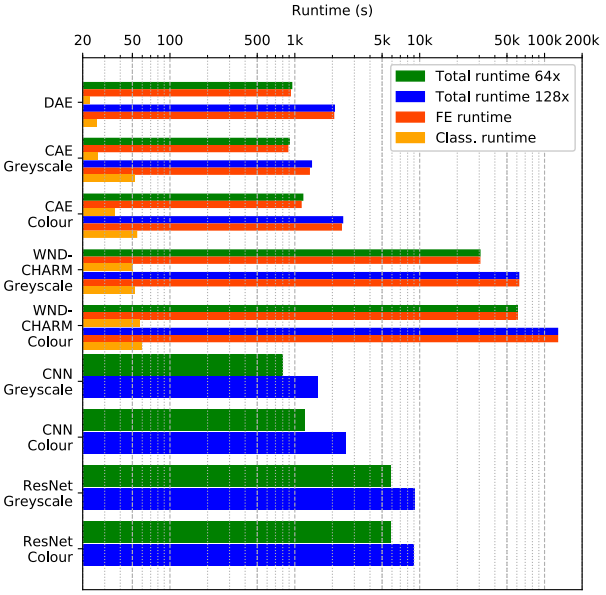


FIGURE 8. Total classification runtime in logarithmic scale for both strategies studied with the GZ1-E sample. For FE approaches, the classification is performed with the RF classifier and embedded FS. This runtime is subdivided in FE runtime and classification runtime.

reconstructions performed for a selection of images from the GZ1-E sample. As it is shown, the DAE model disregards colour channels and is less sensitive to the presence of artefacts in the image and galaxy contours. Conversely, the CAE model defines borders more accurately and partially replicates colour in the images.

These feature sets were taken as input to the classifiers selected for the study. In first place, we carried out the classifications with no FS, pursuing a first comparison of the entire AEs and WND-CHARM feature sets for greyscale and colour images and both image sizes proposed. We then investigated the application of the two FS methods proposed aiming to speed up the classification phase. Accuracy results for these experiments are shown in Table 7 for both image sizes, and comparative representations of the runtime are presented in Figures 6 and 7 for 64x and 128x images, respectively. These values correspond to the average of the classification runtime over the five data partitions.

In third place, we performed the classification of the GZ1-E images using the proposed CNN and the ResNet model as a comparative algorithm. Here we also explored the use of greyscale and colour images and the two image sizes established for the study. These results are shown in Table 8.

Finally, we carried out a comparison among the total classification time of both strategies analysed using the GZ1-E subset. As an estimation of this time for the approaches with FE, we added the FE time of both AE models and WND-CHARM to the classification times obtained for each image configuration (Figures 4, 6 and 7). For the sake of simplicity, here we selected the tandem Embedded FS plus RF classifier for both AE models and WND-CHARM feature sets classification, since this setting offered the best accuracy/runtime trade-off in the experiments presented above (Table 7 and Figures 6 and 7). This is represented in Figure 8.

For a visual illustration of the classification problem investigated and how the different approaches and algorithms work, Figure 9 displays a wide-ranging selection of images from the GZ1-E subset. For simplicity, we restrict the illustration to 128x colour images and select as well the combination Embedded FS plus RF classifier for both AE models and WND-CHARM feature sets classification. We include the result of DAE, CAE, and WND-CHARM FE methods, and both CNN and ResNet classifiers. We also indicate the

TABLE 7. Accuracy results for 64x and 128x images of GZ1-E sample, with no FS (top sector of the table), Fisher scores FS (middle sector) and embedded FS (bottom sector).

		kNN		RF		SVM	
	FE method	64x	128x	64x	128x	64x	128x
No FS	**DAE**	0.8629	0.8648	0.8889	0.8898	0.8743	0.8699
	CAE Greyscale	0.8744	0.8677	0.9127	0.9128	0.8762	0.7685
	CAE Colour	0.8758	0.8742	0.9070	0.9253	0.8353	0.8540
	WND-CHARM Greyscale	0.8848	0.8822	0.9281	0.9255	0.9412	0.9379
	WND-CHARM Colour	0.9184	0.9161	0.9442	0.9444	0.9485	0.9482
Fisher FS	**DAE**	0.8639	0.8651	0.8776	0.8798	0.8765	0.8845
	CAE Greyscale	0.8737	0.8688	0.8505	0.8483	0.8695	0.7733
	CAE Colour	0.8751	0.8734	0.8352	0.8625	0.8299	0.8451
	WND-CHARM Greyscale	0.8816	0.8831	0.9265	0.9255	0.9408	0.9380
	WND-CHARM Colour	0.9193	0.9165	0.9439	0.9446	0.9521	0.9512
Embedded FS	**DAE**	0.8659	0.8648	0.8779	0.8767	0.8412	0.8789
	CAE Greyscale	0.8803	0.8824	0.8516	0.8418	0.8723	0.7120
	CAE Colour	0.8834	0.8914	0.8205	0.8604	0.8339	0.8609
	WND-CHARM Greyscale	0.8917	0.8945	0.9344	0.9314	0.9338	0.9318
	WND-CHARM Colour	0.9201	0.9203	0.9495	0.9490	0.9479	0.9475

TABLE 8. Results of proposed CNN and ResNet for 64x and 128x images of GZ1-E sample.

	64x images		128x images	
	Acc	Runtime (s)	Acc	Runtime (s)
CNN Grey	0.9360	806.9	0.9381	1,539.5
ResNet Grey	0.9181	5,923.3	0.9472	9,136.6
CNN Colour	0.9483	1,207.6	0.9516	2,544.5
ResNet Colour	0.9584	5,879.2	0.9633	9,045.7

amateur label for the object shown, considering the expert classification as ground truth.

B. GZ1-A: AMATEUR SUBSET

Using the GZ1-A subset, the aim of this subsection is two-fold. First, we used this bigger dataset to analyse the scalability of the methods compared in the previous section (Subsection V-B.1). Second, we exploited the huge number of existing amateur-labelled galaxy images by pre-training both CNN and ResNet models on this dataset (Subsection V-B.2).

1) SCALABILITY OF METHODS

After the first set of experiments using the GZ1-E subset, we extended the comparative study to the GZ1-A sample. For this larger dataset, we only compared the features obtained with the CAE and the WND-CHARM for 64x colour images. We completed the classification with RF algorithm, which showed the best balance between runtime and accuracy in the previous experiments with the GZ1-E sample. We also examined the application of both FS methods proposed. These results are shown in Table 9. The representation of classification runtime is presented in Figure 10, which also includes the FE runtime for the CAE and WND-CHARM.

In line with the previous study of GZ1-E, we also carried out the classification with the proposed CNN and ResNet on this sample. Results of accuracy and runtime are shown in Table 10.

TABLE 9. Accuracy results for GZ1-A sample. CAE and WND-CHARM feature sets are classified with RF.

FE method	**No FS**	**Fisher FS**	**Embedded FS**
CAE	0.8481	0.8494	0.8456
WND-CHARM	0.8674	0.8696	0.8734

TABLE 10. Results of the CNN proposed and ResNet for GZ1-A sample.

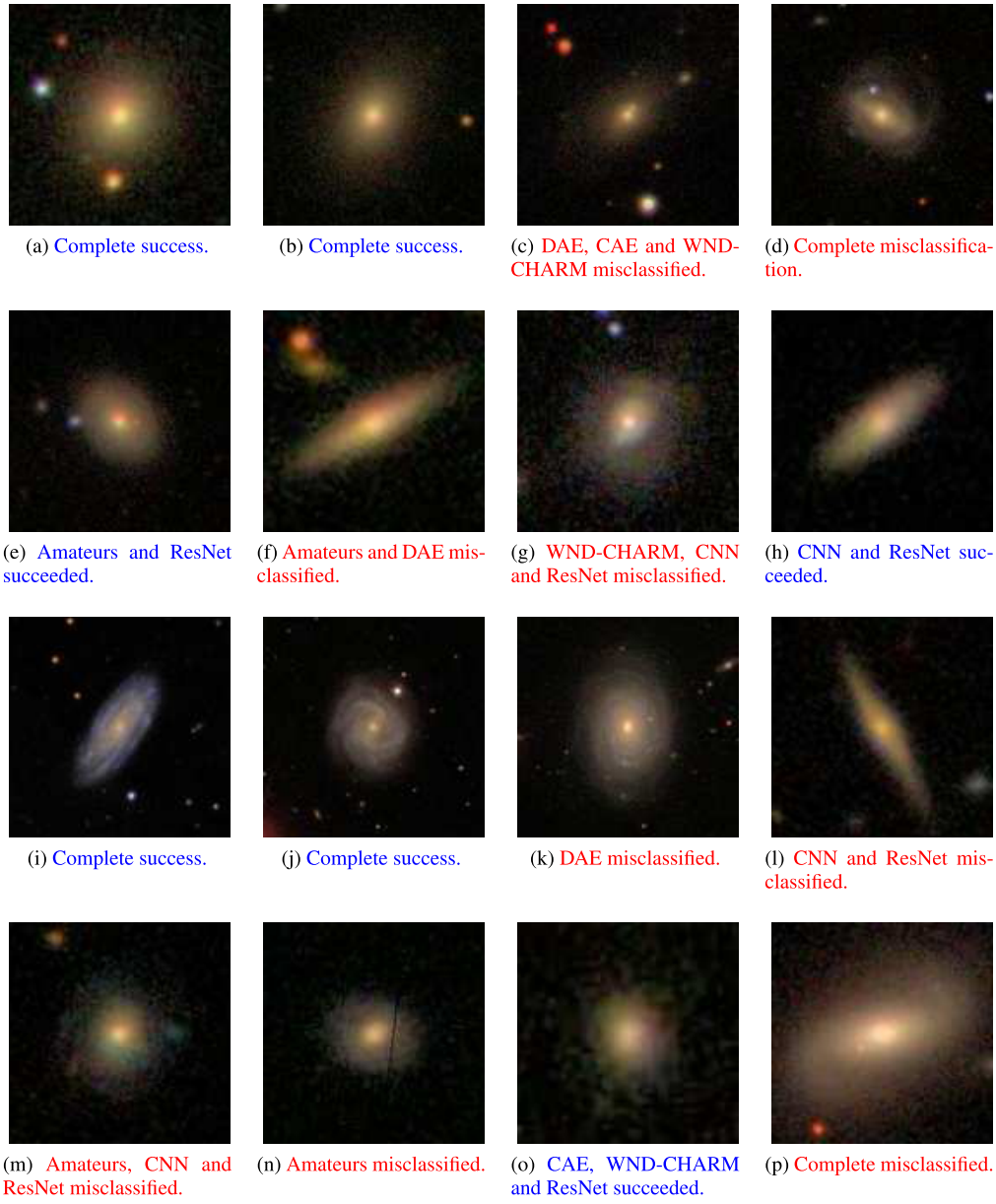
Neural Network	**Acc**	**Runtime** (s)
CNN	0.8989	13,522.3
ResNet	0.9054	96,551.4

Finally, we compared an estimation of the total classification time of both approaches analysed with the GZ1-A sample. As we did with the GZ1-E subset, we added the FE time to the classification runtime for CAE and WND-CHARM. These results are represented in Figure 11.

2) PRE-TRAINING WITH AMATEUR AND EXPERT LABELS

As we explain in Section III-C, citizen science projects enable a novel methodology for the pre-training and fine-tuning of CNNs. By making use of amateur and expert classifications, the network can be pre-trained employing amateur labels, which are expected to be higher in number and coarser in comparison with their expert counterparts. Then, the inclusion of expert labels permits the fine-tuning, occasionally redefining the output (number of classes) of the network.

For this experiment, we considered 64x colour images. We first trained CNN and ResNet on GZ1-A using amateur labels, and then re-trained the network (from previously learned weights for *all* layers) on GZ1-E with expert labels carrying out the usual cross-validation established for all experiments. Since the smaller GZ1-E sample is included in GZ1-A, we removed the overlapping between both samples. That is to say, the pre-train phase skipped the



(a) Complete success. (b) Complete success. (c) DAE, CAE and WND-CHARM misclassified. (d) Complete misclassification.

(e) Amateurs and ResNet succeeded. (f) Amateurs and DAE misclassified. (g) WND-CHARM, CNN and ResNet misclassified. (h) CNN and ResNet succeeded.

(i) Complete success. (j) Complete success. (k) DAE misclassified. (l) CNN and ResNet misclassified.

(m) Amateurs, CNN and ResNet misclassified. (n) Amateurs misclassified. (o) CAE, WND-CHARM and ResNet succeeded. (p) Complete misclassified.

FIGURE 9. Sample of 128x colour images from the GZ1-E subset showing a wide-range of image qualities and difficulty. Two top rows are classified by experts as *elliptical* and two bottom rows as *spiral*. For each of the images, we indicate which approaches misclassify (red) or correctly predict the label (blue), considering expert classifications as ground truth.

examples in GZ1-A later used in the fine-tuning with GZ1-E. We refer to this data sample as GZ1-A*, which consisted of 617,986 examples. In contrast with usual pre-training approaches [61]–[63], here the number of classes does not change and therefore we kept the dense part of both networks for the fine-tuning. Results for this experiment are shown in Table 11.

C. ANALYSIS OF RESULTS

An examination of the tables and charts presented above allowed us to conclude the following remarks after the experiments:

- Among the FE approaches compared, AEs have demonstrated to perform the extraction of features in a shorter amount of time, as it is shown in Figures 4 and 10. However, the classification with the WND-CHARM feature set provided better accuracy compared to AE features across both GZ1 data samples (Table 7 for GZ1-E and 9 for GZ1-A). In broad terms, RF generally outperformed the kNN and SVM algorithms and the use of 64x or 128x images did not make a big difference. Nonetheless, the presence of colour provided better accuracy with respect to greyscale images. This was accomplished at the cost of higher runtime as well, in particular for

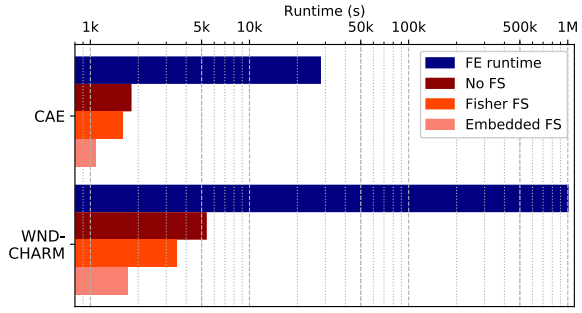


FIGURE 10. FE and classification runtimes in logarithmic scale for GZ1-A sample.

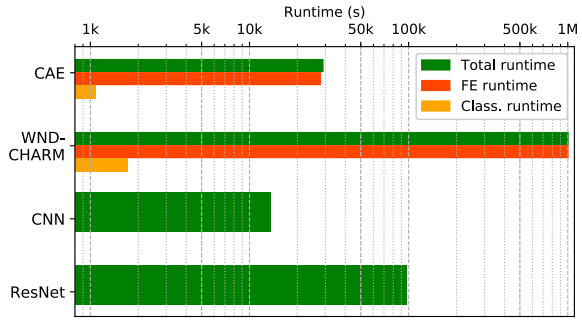


FIGURE 11. Total classification runtime in logarithmic scale for both strategies studied with the GZ1-A sample. For FE approaches, the classification is performed with the RF classifier and embedded FS. This runtime is subdivided in FE runtime and classification runtime.

TABLE 11. Results for the proposed CNN and ResNet, implementing the pre-training and fine-tuning with the GZ1-A* and GZ1-E subsets, respectively.

Neural Network	Acc	Runtime **Pre-training** (s)	Runtime **Fine-tuning** (s)
CNN	0.9574	55,419.3	841.8
ResNet	0.9643	107,430.63	5,775.0

the WND-CHARM Colour feature extractor, which has proved to be the best feature set in terms of classification accuracy.

- The two FS methods proposed did not have a big impact on the classification accuracy. However, they considerably diminished the classification runtime, especially for the WND-CHARM feature sets, as it is presented in Figures 6 and 7. The most promising results were obtained for the embedded FS method, probably due to the dynamic nature of this approach that selects a variable number of features. Conversely, the method based on Fisher scores always filters a fixed number of features, thus providing a runtime reduction that remained steady.
- Although the WND-CHARM Colour feature set yielded better classification accuracy, both AE models proposed here have proved to greatly accelerate the FE process (Figures 4 and 10), which could be decisive for the classification of big volumes of data. Among the AE models proposed, the CAE provided the best results in terms of accuracy with respect to the DAE, also enabling the use of colour images. For this architecture, the global classification time with the GZ1-E dataset is comparable to the classification using the proposed CNN (Figure 8). This confirms the potential utility of AEs in the classification of large amounts of image data, given that the AE's training would be completed only once.
- Both analysed CNNs provided the best performance in comparison with the three FE approaches studied in terms of accuracy/runtime balance for both data samples used in the study (Tables 8 and 10). Nonetheless, the WND-CHARM Colour feature set was able to obtain comparable accuracy in the GZ1-E subset when the classification was made using the SVM algorithm (Table 7). The difference was enlarged with the GZ1-A sample (Table 9), showing that CNNs coped better with the learning from larger amounts of data that probably contain more noise in the labels, and also revealing that amateur labels tended to degrade the classification accuracy.
- ResNet generally outperformed the proposed CNN model in terms of accuracy in the experiments with GZ1-E with expert labels. However, this was achieved at the cost of a much higher runtime that grew up to seven times for greyscale images (Table 8). In contrast, the improvement was marginal in the classification of GZ1-A with amateur labels, where a huge increase of the runtime did not provide a much better accuracy (Table 10). These experiments reveal that deeper architectures do not always translate in much more improved results, and that the selection of the model can be critical for an optimal DL classification approach in terms of time.
- The proposed pre-training and fine-tuning scheme, using both amateur and expert labels, showed a promising result as a way of leveraging all the potential of citizen science projects (Table 11). Here the improvement in accuracy (comparing with the previous experiment with GZ1-E only, Table 8), was greater for the proposed CNN, indicating that the addition of more layers in the network did not provide any substantial improvement with coarser labels. However, the pre-training phase considerably enlarged the total runtime for both CNNs. This experiment confirmed the adequacy of considering expert and amateur labels to feed the learning of a ML approach, which specially applies in complex classification problems such as the one studied in this paper. For example, images 9d, 9e, 9l and 9n demonstrate that even the best approaches compared in the study are prone to fail (if we consider experts' judgements as ground truth), and an integrated use of all knowledge about the problem (e.g. expert classifications, additional astronomical data, citizen science results) can be crucial.

VI. CONCLUSION

In this paper, we have presented a comparative study about the performance of two different strategies for the automated classification of galaxy images, either classifying a feature

set obtained from the image or with convolutional neural networks. Through a set of experiments, we have compared the state-of-the-art feature extractor, the WND-CHARM, with the suitability of autoencoders for feature extraction of galaxy images. We have then compared these results with the end-to-end classification provided by two models of convolutional neural networks under the same experimental setting. We have explored the impact of the image size and the presence or not of the colour channels in the classification results, also studying the effect of two distinct feature selection methods. The experiments have been run using two different samples from the Galaxy Zoo 1 image dataset, also studying the scalability of both approaches to larger data and the influence of amateur and expert classifications in the classification accuracy. In addition, we have introduced a novel approach based on pre-training and fine-tuning of convolutional neural networks that have proven to take advantage of both label sets available for this problem.

The results allow us to conclude that convolutional neural networks offer the best trade-off between runtime and accuracy although the addition of a big depth and complexity in the network does not always provide a significant improvement in their prediction capability, depending on the classification problem at hand. Also, autoencoders represent a promising alternative for the classification of these images with feature extraction. This is a consequence of their ability to separate the feature extraction and learning processes, which could eventually be beneficial when the amount of data to be classified expands. Finally, it has been shown that very promising results may eventually come from the learning of both amateur and expert label sets that citizen science projects offer. Following the work presented here, we plan to enhance the learning phase with the consideration of unlabelled data in conjunction with different levels of confidence in the images labelling.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, no. 1, pp. 249–268, 2007.
[2] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann, 2005.
[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2001.
[4] D. Peralta, I. Triguero, S. García, Y. Saeys, J. M. Benitez, and F. Herrera, "On the use of convolutional neural networks for robust classification of multiple fingerprint captures," *Int. J. Intell. Syst.*, vol. 33, no. 1, pp. 213–230, Nov. 2018.
[5] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognit. Lett.*, vol. 120, pp. 69–74, Apr. 2019.
[6] D. Lu, M. Heisler, S. Lee, G. W. Ding, E. Navajas, M. V. Sarunic, and M. F. Beg, "Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network," *Med. Image Anal.*, vol. 54, pp. 100–110, May 2019.
[7] M. Umehara, H. S. Stein, D. Guevarra, P. F. Newhouse, D. A. Boyd, and J. M. Gregoire, "Analyzing machine learning models to accelerate generation of fundamental materials insights," *npj Comput. Mater.*, vol. 5, no. 1, pp. 1–9, Mar. 2019.
[8] J. Tyson, "Optical synoptic telescopes: New science frontiers," in *Proc. The Int. Soc. for Opt. Eng.*, vol. 7733, no. 1, 2010, Art. no. 773303.
[9] K. D. Borne, "Astroinformatics: Data-oriented astronomy research and education," *Earth Sci. Informat.*, vol. 3, nos. 1–2, pp. 5–17, May 2010.
[10] S. Ackermann, K. Schawinski, C. Zhang, A. K. Weigel, and M. D. Turp, "Using transfer learning to detect galaxy mergers," *Monthly Notices Roy. Astronomical Soc.*, vol. 479, no. 1, pp. 415–425, May 2018.
[11] R. E. González, R. P. Muñoz, and C. A. Hernández, "Galaxy detection and identification using deep learning and data augmentation," *Astron. Comput.*, vol. 25, pp. 103–109, Oct. 2018.
[12] C. E. Petrillo, C. Tortora, S. Chatterjee, G. Vernardos, L. V. E. Koopmans, G. V. Kleijn, N. R. Napolitano, G. Covone, L. S. Kelvin, and A. M. Hopkins, "Testing convolutional neural networks for finding strong gravitational lenses in KiDS," *Monthly Notices Roy. Astronomical Soc.*, vol. 482, no. 1, pp. 807–820, 2019.
[13] E. Hubble, "Extra-galactic nebulae," *Astrophys. J.*, vol. 64, pp. 321–373, Dec. 1926.
[14] A. Sandage,"The classification of galaxies: Early history and ongoing developments," *Annu. Rev. Astron. Astrophys.*, vol. 43, no. 1, pp. 581–624, Sep. 2005.
[15] K. Borne, *Virtual Observatories, Data Mining, and Astroinformatics*. Dordrecht, The Netherlands: Springer, 2013.
[16] J. P. Cohn, "Citizen science: Can volunteers do real research?" *BioScience*, vol. 58, no. 3, pp. 192–197, Mar. 2008.
[17] R. Simpson, K. Page, and D. De Roure, "Zooniverse: Observing the world's largest citizen science platform," in *Proc. Int. Conf. World Wide Web*, 2014, pp. 1049–1054.
[18] C. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. Raddick, R. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, "Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey," *Monthly Notices Roy. Astronomical Soc.*, vol. 389, no. 3, pp. 1179–1189, 2008.
[19] C. Lintott, K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, E. Edmondson, K. Masters, R. Nichol, M. Raddick, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, "Galaxy Zoo 1: Data release of morphological classifications for nearly 900 000 galaxies," *Monthly Notices Roy. Astronomical Soc.*, vol. 410, no. 1, pp. 166–178, 2011.
[20] Z. Ivezić, S. Kahn, J. Tyson, B. Abel, E. Acosta, R. Allsman, D. Alonso, Y. Alsayyad, S. Anderson, J. Andrew, and, "LSST: From science drivers to reference design and anticipated data products," *Astrophysical J.*, vol. 873, no. 2, pp. 1–44, 2019.
[21] M. Banerji, O. Lahav, C. Lintott, F. Abdalla, K. Schawinski, S. Bamford, D. Andreescu, P. Murray, M. Raddick, A. Slosar, A. Szalay, D. Thomas, and J. Vandenberg, "Galaxy zoo: Reproducing galaxy morphologies via machine learning," *Monthly Notices Roy. Astronomical Soc.*, vol. 406, no. 1, pp. 342–353, 2010.
[22] L. Shamir, "Automatic detection of peculiar galaxies in large datasets of galaxy images," *J. Comput. Sci.*, vol. 3, no. 3, pp. 181–189, May 2012.
[23] S. Dieleman, K. W. Willett, and J. Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction," *Monthly Notices Roy. Astronomical Soc.*, vol. 450, no. 2, pp. 1441–1459, Apr. 2015.
[24] X.-P. Zhu, J.-M. Dai, C.-J. Bian, Y. Chen, S. Chen, and C. Hu, "Galaxy morphology classification with deep convolutional neural networks," *Astrophys. Space Sci.*, vol. 364, no. 4, pp. 1–15, Apr. 2019.
[25] M. Nixon and A. Aguado, *Feature Extraction & Image Processing*, 2nd ed. Orlando, FL, USA: Academic, 2008.
[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[27] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[28] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.
[29] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[32] D. Charte, F. Charte, S. García, M. J. del Jesus, and F. Herrera, "A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines," *Inf. Fusion*, vol. 44, pp. 78–96, Nov. 2018.

[33] F. J. Pulgar, F. Charte, A. J. Rivera, and M. J. del Jesus, "AEkNN: An AutoEncoder kNN-based classifier with built-in dimensionality reduction," *Int. J. Comput. Intell. Syst.*, vol. 12, no. 1, pp. 436–452, 2018.

[34] N. Ball and R. Brunner, "Data mining and machine learning in astronomy," *Int. J. Mod. Phys. D*, vol. 19, no. 7, pp. 1049–1106, 2010.

[35] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, and I. G. Goldberg, "WND-CHARM: Multi-purpose image classification using compound image transforms," *Pattern Recognit. Lett.*, vol. 29, no. 11, pp. 1684–1693, Aug. 2008.

[36] L. Shamir, A. Holincheck, and J. Wallin, "Automatic quantitative morphological analysis of interacting galaxies," *Astron. Comput.*, vol. 2, pp. 67–73, Aug. 2013.

[37] E. Kuminski and L. Shamir, "A computer-generated visual morphology catalog of ~ 3,000,000 SDSS galaxies," *Astrophys. J. Suppl. Ser.*, vol. 223, no. 2, pp. 1–10, 2016.

[38] S. van den Bergh, "A new classification system for galaxies," *Astrophys. J.*, vol. 206, pp. 883–887, Jun. 1976.

[39] R. E. Hart, S. P. Bamford, W. C. Keel, S. J. Kruk, K. L. Masters, B. D. Simmons, and R. J. Smethurst, "Galaxy Zoo: Constraining the origin of spiral arms," *Monthly Notices Roy. Astronomical Soc.*, vol. 478, no. 1, pp. 932–949, May 2018.

[40] C. Watson, K.-V. Tran, A. Tomczak, L. Alcorn, I. V. Salazar, A. Gupta, I. Momcheva, C. Papovich, P. V. Dokkum, G. Brammer, J. Lotz, and C. N. A. Willmer, "Galaxy merger fractions in two clusters at $z \sim 2$ using the hubble space telescope," *Astrophys. J.*, vol. 874, no. 1, p. 63, Mar. 2019.

[41] R. M. G. Delgado, E. Pérez, R. C. Fernandes, R. García-Benito, R. L. Fernández, N. V. Asari, C. Cortijo-Ferrero, A. L. de Amorim, E. A. D. Lacerda, S. F. Sánchez, M. D. Lehnert, and C. J. Walcher, "Spatially-resolved star formation histories of CALIFA galaxies: Implications for galaxy formation," *Astron. Astrophys.*, vol. 607, no. 128, pp. 1–21, 2017.

[42] R. Bonney, J. L. Shirk, T. B. Phillips, A. Wiggins, H. L. Ballard, A. J. Miller-Rushing, and J. K. Parrish, "Next steps for citizen science," *Science*, vol. 343, no. 6178, pp. 1436–1437, Mar. 2014.

[43] L. Trouille, C. J. Lintott, and L. F. Fortson, "Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 6, pp. 1902–1909, Feb. 2019.

[44] M. Jiménez, I. Triguero, and R. John, "Handling uncertainty in citizen science data: Towards an improved amateur-based large-scale classification," *Inf. Sci.*, vol. 479, pp. 301–320, Apr. 2019.

[45] K. Willett, C. Lintott, S. Bamford, K. Masters, B. Simmons, K. Casteels, E. Edmondson, L. Fortson, S. Kaviraj, W. Keel, T. Melvin, R. Nichol, M. J. Raddick, K. Schawinski, R. Simpson, R. Skibba, A. Smith, and D. Thomas, "Galaxy Zoo 2: Detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey," *Monthly Notices Roy. Astronomical Soc.*, vol. 435, no. 4, pp. 2835–2860, 2013.

[46] K. Willett, M. A. Galloway, S. P. Bamford, C. J. Lintott, K. L. Masters, C. Scarlata, B. D. Simmons, M. Beck, C. N. Cardamone, E. Cheung, and E. M. Edmondson, "Galaxy Zoo: Morphological classifications for 120 000 galaxies in *HST* legacy imaging," *Monthly Notices Roy. Astronomical Soc.*, vol. 464, no. 4, pp. 4176–4203, 2017.

[47] M. Thonnat and M. Berthod, "Automatic classification of galaxies into morphological types," in *Proc. Int. Conf. Pattern Recognit.*, vol. 2, 1984, pp. 844–846.

[48] O. Lahav, A. Naim, R. J. Buta, H. G. Corwin, G. de Vaucouleurs, A. Dressler, J. P. Huchra, S. van den Bergh, S. Raychaudhury, L. Sodre, and M. C. Storrie-Lombardi, "Galaxies, human eyes, and artificial neural networks," *Science*, vol. 267, no. 5199, pp. 859–862, Feb. 1995.

[49] I. A. Basheer and M. Hajmeer, "Artificial neural networks: Fundamentals, computing, design, and application," *J. Microbiol. Methods*, vol. 43, no. 1, pp. 3–31, Dec. 2000.

[50] C. J. Conselice, M. A. Bershady, and A. Jangren, "The asymmetry of galaxies: Physical morphology for nearby and high-redshift galaxies," *Astrophys. J.*, vol. 529, no. 2, pp. 886–910, Feb. 2000.

[51] J. M. Lotz, J. Primack, and P. Madau, "A new nonparametric approach to galaxy morphological classification," *Astronomical J.*, vol. 128, no. 1, pp. 163–182, Jul. 2004.

[52] J. De La Calleja and O. Fuentes, "Machine learning and image analysis for morphological galaxy classification," *Monthly Notices Roy. Astronomical Soc.*, vol. 349, no. 1, pp. 87–93, Mar. 2004.

[53] M. Huertas-Company, D. Rouan, L. Tasca, G. Soucail, and O. Le Fèvre, "A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images: I. method description," *Astron. Astrophys.*, vol. 478, no. 3, pp. 971–980, Nov. 2008.

[54] L. Shamir, N. Orlov, D. M. Eckley, T. Macura, J. Johnston, and I. G. Goldberg, "Wndchrm—An open source utility for biological image analysis," *Source Code for Biol. Med.*, vol. 3, no. 1, pp. 1–13, Jul. 2008.

[55] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[57] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[58] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," 2019, *arXiv:1908.07919*. [Online]. Available: http://arxiv.org/abs/1908.07919

[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[60] D. Yu and M. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 237–240.

[61] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning Earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[62] U. K. Lopes and J. F. Valiati, "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection," *Comput. Biol. Med.*, vol. 89, pp. 135–143, Oct. 2017.

[63] B. Kieffer, M. Babaie, S. Kalra, and H. R. Tizhoosh, "Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks," in *Proc. 7th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2018, pp. 1–6.

[64] M. Huertas-Company, R. Gravet, G. Cabrera-Vives, P. G. Pérez-González, J. S. Kartaltepe, G. Barro, M. Bernardi, S. Mei, F. Shankar, P. Dimauro, E. F. Bell, D. Kocevski, D. C. Koo, S. M. Faber, and D. H. Mcintosh, "A catalog of visual-like morphologies in the 5 CANDELS fields using deep learning," *Astrophys. J. Suppl. Ser.*, vol. 221, no. 1, pp. 1–23, Oct. 2015.

[65] M. Ranzato, Y. Boureau, Y. LeCun, and S. Chopra, "A unified energy-based framework for unsupervised learning," *J. Mach. Learn. Res.*, vol. 2, pp. 371–379, 2007.

[66] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.

[67] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[68] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[69] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[70] M. Mahdianpari, B. Salehi, M. Rezaee, F. Mohammadimanesh, and Y. Zhang, "Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery," *Remote Sens.*, vol. 10, no. 7, p. 1119, Jul. 2018.

[71] F. Taherkhani, H. Kazemi, A. Dabouei, J. Dawson, and N. Nasrabadi, "A weakly supervised fine label classifier enhanced by coarse supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6459–6468.

[72] X. Wu, V. Kumar, Q. Ross, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[73] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[74] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[75] C. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.
[76] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Aug. 2007.
[77] R. G. J. Wijnhoven and P. H. N. de With, "Fast training of object detection using stochastic gradient descent," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 424–427.
[78] S. Akila Agnes and J. Anitha, "Analyzing the effect of optimization strategies in deep convolutional neural network," in *Nature Inspired Optimization Techniques for Image Processing Applications* (Intelligent Systems Reference Library), vol. 150. Cham, Switzerland: Springer, 2019, pp. 235–253.

MANUEL JIMÉNEZ received the M.Sc. degree in physics from the University of Granada, Granada, Spain, in 2016. He is currently pursuing the Ph.D. degree with the Computational Optimisation and Learning (COL) Laboratory, School of Computer Science, University of Nottingham. His research interests include the application of machine learning techniques to the classification of astronomical images under conditions of uncertainty.

MERCEDES TORRES TORRES received the M.Sc. degree in digital signal and image processing from Cranfield University, in 2010, and the Ph.D. degree in computer science from the University of Nottingham, in 2014. She has been an Assistant Professor of computer science with the University of Nottingham, since 2015, where she is a member of the Computer Vision Laboratory and the Horizon Research Institute. She works in the development and application of new machine learning techniques, particularly deep learning, for skewed or small datasets. She is currently leading the Knowledge Transfer Partnership funded by the Innovate U.K., and Netacea investigating fine-grained web traffic classification.

ROBERT JOHN (Senior Member, IEEE) received the bachelor's degree (Hons.) in mathematics, the M.Sc. degree in statistics, and the Ph.D. degree in type-2 fuzzy logic. His initial career was as a mathematician in various roles for industry and commerce as an AI Consultant. He joined De Montfort University, in 1989, and the University of Nottingham, in 2013. He is currently a Professor of operational research and computer science, and the Head of the Computational Optimisation and Learning (COL) Laboratory, School of Computer Science, University of Nottingham. He is also an elected member of the EPSRC College. In the field of type-2 fuzzy logic, his work is widely recognized by the international fuzzy logic community as leading in the aspects of theoretical foundations and practical applications. His work has produced many fundamental new results that have opened the field to new research, enabling a broadening of scope and application. He has significant funding for his research from a variety of sources. He has over 8500 citations on Google Scholar with an H-index of 39. He is also a Fellow of the British Computer Society.

ISAAC TRIGUERO (Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Granada, Spain, in 2009 and 2014, respectively. He has been an Assistant Professor of data science with the University of Nottingham, since June 2016. His work is mostly concerned with the research of novel methodologies for big data analytics. He has published more than 70 international publications in the fields of big data, machine learning, and optimization (H-index = 24 and more than 2300 citations on Google Scholar). He has acted as the Program Co-Chair of the IEEE Conference on Smart Data, in 2016, the IEEE Conference on Big Data Science and Engineering, in 2017, and the IEEE International Congress on Big Data, in 2018. He is currently leading the Knowledge Transfer Partnership Project funded by the Innovative U.K. and the energy provider E.ON, that investigates smart metering data. He is also the Section Editor-in-Chief of the *Machine Learning and Knowledge Extraction* journal and an Associate Editor of the *Big Data and Cognitive Computing* journal and IEEE Access journal.