IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received January 24, 2022, accepted February 8, 2022, date of publication February 10, 2022, date of current version February 24, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150881

Transfer Learning and Deep Metric Learning for Automated Galaxy Morphology Representation

MOHAMED ZAYYAN VARIAWA[1], TERENCE L. VAN ZYL[1], (Member, IEEE), AND MATTHEW WOOLWAY[2]

[1]Institute for Intelligent Systems, University of Johannesburg, Johannesburg 2006, South Africa
[2]Faculty of Engineering and the Built Environment, University of Johannesburg, Johannesburg 2006, South Africa

Corresponding author: Matthew Woolway (matt.woolway@gmail.com)

This work was supported in part by Nedbank Research Chair.

ABSTRACT Galaxy morphology characterisation is an important area of study, as the type and formation of galaxies offer insights into the origin and evolution of the universe. Owing to the increased availability of images of galaxies, scientists have turned to crowd-sourcing to automate the process of instance labelling. However, research has shown that using crowd-sourced labels for galaxy classification comes with many pitfalls. An alternative approach to galaxy classification is metric learning. Metric learning allows for improved representations for classification, anomaly detection, information retrieval, clustering and dimensionality reduction. Understanding the implications of this approach regarding crowd-sourced labels is of paramount importance if scientists intend to continue using them. This paper compares metric learning and classification models trained or fine-tuned on both the crowd-sourced Galaxy Zoo 2 (GZ2-H) dataset and expertly labelled EFIGI catalogue. The study uses the Revised Shapley-Ames (RSA) catalogue of bright galaxies, also labelled by experts, as an unseen test set. The RSA catalogue allows for an accurate comparison of the performance of the models at predicting the Hubble types of galaxies. The classification accuracy for the crowd-sourced and expert models indicated that the models are comparable on the surface. However, using alternative metrics, the results show that the models trained on the expert dataset outperformed the model trained on the crowd-sourced data in terms of actual vs predicted labels. Further, the results show that fine-tuning a model pre-trained on crowd-sourced data can outperform the state-of-the-art in galaxy characterisation. The models trained to predict the Hubble types of galaxies are better when fine-tuned using the Proxy-NCA and Normalised-Softmax loss functions than with other pairwise losses. The Normalised-Softmax loss yielded the best overall 9-class models with accuracies at 30.88% (GZ2-H) and 30.05% (EFIGI) and MAP values of 0.3483 (GZ2-H) and 0.3889. The Proxy-NCA loss produced the second-best overall 9-class models with accuracies at 30.33% (GZ2-H) and 20.03% (EFIGI) and MAP values of 0.3577 (GZ2-H) and 0.3917 (EFIGI). Finally, the paper highlights the need for caution when utilising crowd-sourced labels; however, it argues that transfer learning from crowd-sourced labelled data to expert-labelled data can still lead to significant improvements.

INDEX TERMS Deep metric learning, galaxy classification, transfer learning, crowd-sourced labels.

NOMENCLATURE

DATASET

GZ1	Galaxy Zoo 1.
GZ2	Galaxy Zoo 2.
GZ2-H	Galaxy Zoo 2 Hubble types.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

LOSS

CE	Cross-entropy.
Contr	Contrastive.
LS-CE	Label Smoothing Cross-entropy.
Norm-S'max	Normalised Softmax.
P-anc	Proxy Anchor.
P-NCA	Proxy NCA.

MODEL

Res50	ResNet50.

TYPE

Class	Classification.

I. INTRODUCTION

In trying to understand the origin and evolution of the universe, an important area of focus is galaxy characterisation. The type and formation of galaxies offer clues and insights into the development of the universe [1]. Galaxy morphological characterisation separates galaxies into classes based on their physical structures. A galaxies' morphological characterisation typically falls within three major categories; namely elliptical, spiral and irregular. An elliptical galaxy has an ellipse-shaped light profile. Spiral galaxies are disk-shaped and have multiple curved arms originating from the centre. Irregular galaxies do not fit into the elliptical and spiral categories. A widely accepted system for this characterisation of galaxies is the Hubble tuning fork [2].

The majority of labels available in catalogues come from experts inspecting images of galaxies and manually assigning them. The significant increase in the amount of data being made available (in the order of $1e^6$) has made it increasingly laborious for scientists to classify these galaxies manually [2]. One approach is to use machine learning to automate this task [3]. However, the number of expertly-labelled galaxies in some catalogues such as the EFIGI [4] and Revised Shapely-Ames (RSA) [5] are severely limited at 4488 and 1249 samples, respectively. This scarcity of expertly-labelled examples would be detrimental to the success of this classification task in a deep learning context.

A consequence of this lack of labelled images and a shortage of experts is citizen-science becoming an increasingly popular approach for assigning labels to images of galaxies [2], [3], [6], [7]. For example, Galaxy Zoo is a widely popular citizen-science, galaxy classification project to reduce the time spent by astronomers manually labelling images of galaxies [3]. Considering how one might combine the advances in machine learning and deep learning with the advantages of citizen-science in overcoming these challenges is an active area of research [3], [8]–[10].

If researchers are to couple citizen-science and deep learning to automate galaxy characterisation, it is essential to understand how well these crowd-sourced labels generalise to expertly labelled catalogues. Previously, Variawa *et al.* [11] trained a model on the 37-class response vectors in the Galaxy Zoo 2 dataset as a base model for two experiments. The first experiment fine-tuned the model to predict the Hubble types of galaxies in the Galaxy Zoo 2 dataset. The second experiment fine-tuned a model to predict the Hubble types of galaxies in the EFIGI catalogue. They used the RSA catalogue as an additional, expertly labelled held out test set to compare the generalisation between crowd-sourced and expertly labelled data. Their results show that despite being able to achieve state-of-the-art classification performance when measured against a test set split from that catalogue, the models did not generalise well to the unseen RSA catalogue [11].

The research above raises severe concerns around the efficacy of galaxy characterisation using a classification of crowd-sourced labels with deep learning. An alternate approach to classification is using metric learning to ascertain the similarity between objects [12]–[18]. Metric Learning learns a function that maps objects into a representational embedding space. The aim is for this representation to preserve the ‘‘distance’’ between objects - similar samples are closer to each other, and dissimilar ones are further apart from one another [19], [20]. Metric learning also has applications in image recognition and multi-label classification [21]–[23] and has the potential to be applied to other astronomical use cases [24]–[26].

Since metric learning has fewer constraints than classification and potential additional applications to unsupervised learning tasks like information retrieval and clustering, it is worth exploring if metric learning provides any benefits when tested on the unseen RSA catalogue.

This paper reports on experiments to test the generalisation of the Galaxy Zoo 2 crowd-sourced data and EFIGI expert labelled data to the RSA catalogue. The RSA catalogue served as a test set, as in previous research. The RSA catalogue allowed us to contrast models on a completely unseen, expertly labelled set. First, the results show improvement on the state-of-the-art in galaxy classification [11] using deep metric learning techniques [16]. These results are evidenced by improved accuracy not equating to improved, learned representations. Second, the results demonstrate that a model trained using Label Smoothing with the Cross-Entropy loss can be fine-tuned using deep metric learning to enhance the models' ability to predict the Hubble types of galaxies. Third, we present evidence that deep learning models fine-tuned on the expertly labelled catalogue are better at predicting the Hubble types of galaxies than deep learning models fine-tuned on crowd-sourced data. Finally, the results show that transfer learning from crowd-sourced data to expert labelled data achieves state-of-art results.

The research presented here provides evidence and support for an alternative methodology for using crowd-sourced labels for galaxy characterisation. Further, this research has shown that deep metric learning can improve upon the current state but does not necessarily in and of itself overcome the limitations of crowd-sourced data.

A. THE GALAXY ZOO PROJECT

Galaxy Zoo is a citizen-science, galaxy classification project described by Lintott *et al.* [3]. The objective of the project was to reduce the time spent by astronomers having to classify galaxies manually. Each image came with a set of questions about the structure of the galaxy in the image [3]. There are 11 questions, each with a pre-defined set of responses resulting in 37 possible responses.

A Kaggle competition centred around the Galaxy Zoo project supplied participants with 61578 images from the Galaxy Zoo project to train machine learning models to predict the 37-class response vectors. Dieleman *et al.* [27] produced the winning model, a Convolutional Neural Network (CNN). The model achieved a Root Mean Squared Error (RMSE) of 0.0747 on the unseen test set. The pre-processing step involved cropping and down-scaling the images followed by the use of data augmentation techniques (e.g. random rotations and flipping, randomly re-scaling the size and adjusting the brightness) to transform the images.

Lukic and Bruggen [28] replicated the Galaxy Zoo Kaggle competition's winning solution by training a CNN with varying convolutional and pooling layers. The experiments used the activation functions of the Rectified Linear Unit (ReLU) and Parametric Rectified Linear Unit (PReLU). Stochastic Gradient Descent (SGD) with a minibatch size of 16 was utilised for training the networks. The models were trained for 200 epochs, with an initial learning rate of $4e^{-2}$, and later reduced to $4e^{-4}$ at epoch 150. Results showed that three convolutional layers and the PReLU activation function produced lower training and validation errors than the ReLU function using the mean squared error (MSE) loss function. Using the PReLu activation function, Lukic and Bruggen [28] reported an RMSE of 0.4111 on the training set and an RMSE of 0.4062 on the validation set. The differences in RMSE values reported by Lukic and Bruggen [28] and Dieleman *et al.* [27] are attributed to the difference in CNN architectures, in addition to the advanced image pre-processing techniques used by Dieleman *et al.* [27].

Following the conclusion of the Galaxy Zoo 1 project, Katebi *et al.* [29] trained a Capsule Network to predict the 37-class response vectors in the Galaxy Zoo 1 dataset. The Capsule Network produced an RMSE of 0.103 on the unseen test set. Katebi *et al.* [29] extended this approach by selecting the answer with the largest number of responses as the sample label and training a Capsule Network classifier with the responses to the first question in the Galaxy Zoo decision tree [3]. The first question, "Is the galaxy simply smooth and rounded with no sign of a disk?" had three responses, namely, "round and smooth" (i.e. elliptical galaxies), "objects with disks" (spiral galaxies) and "artefact or star." These three responses were selected as the target classes for training, with the Capsule Network achieving a classifying accuracy of 98.77%, a modest improvement over the model baseline accuracy of 96.96%.

Variawa *et al.* [30] trained a ResNet50 model to predict the 37-class vectors in the Galaxy Zoo 1 dataset, achieving a RMSE of 0.0942 on the unseen test set. A decision tree, available in the original Galaxy Zoo 1 publication [3] was used to define a set of rules mapping the 37-class response vectors to Hubble types. The model was then used to predict the 37-class vectors for galaxies in the RSA catalogue. The self-defined rules then mapped these vectors to a Hubble type. Transfer learning was employed wherein the model trained to predict the 37-class response vectors served as initialisation to train a model using the RSA catalogue to predict the Hubble labels. Variawa *et al.* [30] reported that neither rules-based nor transfer learning approaches could reliably predict the Hubble types of galaxies in the RSA catalogue [30].

Galaxy Zoo 2 extended the original Galaxy Zoo project by including more information on each galaxies' structure, such as bars, spirals, bulges, and others [31]. The Galaxy Zoo 2 project consisted of 300 thousand galaxies taken from the Sloan Digital Sky Survey (SDSS). Furthermore, Galaxy Zoo 2 included the Hubble type for each galaxy (based on participants' responses) and the 37 responses from the Galaxy Zoo project [8], [31].

Barchi *et al.* [31] trained a Decision Tree, a Support Vector Machine, a Multi-Layer Perceptron and a CNN on the Galaxy Zoo 2 Hubble types. The models were tested on a various classes, namely 3, 7, 9 and 11. For the 3-class classification, the Decision Tree achieved an accuracy of 78.70%, the Support Vector Machine achieved an accuracy of 78.50%, the Multi-Layer Perceptron achieved an accuracy of 78.80%, and the CNN performed the best with an accuracy of 82.70%; however, the accuracies dropped when the number of classes increased. Specifically, the accuracy of the CNN dropped to 70.00% for the 7-class classification, 67.40% for the 9-class classification and 65.20% when classifying galaxies into 11 classes.

Gupta *et al.* [32] present a continuous-depth variation of the ResNet architecture that uses Neural Ordinary Differential Equations (NODE) for galaxy classification on the GZ2 dataset. The advantage of using NODE over the standard ResNet architecture is that it takes less time to train and requires fewer training samples while achieving accuracy comparable to ResNet. However, numerical errors can occur when computing the gradient during back-propagation. The Adaptive Checkpoint Adjoint (ACA) method can mitigate numerical errors when calculating the gradient during back-propagation. The ACA method used in conjunction with NODE is NODE_ACA. Classifying galaxies into five classes, NODE_ACA achieves an average accuracy of 84.2%, while ResNet achieves 77.88%.

Self-supervised learning was employed to deduce representations of images from the Sloan Digital Sky Survey (SDSS). These representations may be used as input features or further fine-tuned to outperform models trained only on labelled data. Hayat *et al.* [24] use the learned representations train for galaxy morphology classification on the GZ2 data. An encoder (a ResNet50 architecture) learns a 2048-dimensional vector (i.e. representation) of the training images. Augmentation techniques transformed the samples before being loaded into the encoder. A contrastive loss ensured that augmentations from the same image had similar representations, while augmentations of different images would have dissimilar representations. Three classification models were trained on a subset of the GZ2 questions, where each question was considered a separate binary classification task. The first classification model was a CNN trained in a supervised learning fashion. The second model was a linear classifier trained on the learned representations, and the third was the self-supervised encoder fine-tuned for a few iterations. The results showed that both the linear classifier and the fine-tuned self-supervised encoder outperformed the CNN with a limited number of labels. Specifically, a factor of 16 more labels is required in the CNN to achieve the same performance as the fine-tuned self-supervised models.

Moonzarin Reza [33] trained five machine learning algorithms on data from the SDSS to classify galaxies into four classes; spirals, ellipticals, mergers and stars (unknown). Classifying the mergers as a separate class presented a challenge as this class could easily be confused with either the spiral or elliptical classes [33]. Principal Components Analysis (PCA) was used to extract the 25 most significant principal components. The machine learning models used these 25 components as the input. Comparing the results with the Galaxy Zoo labels, Moonzarin Reza [33] reported a test accuracy of 98.20% by an Artificial Neural Network and 97.50% by the ExtraTrees classifier. However, when classifying the merger and star classes, the ExtraTrees classifier outperformed the Artificial Neural Network.

Walmsley *et al.* [34] present Galaxy Zoo DECaLS, another phase in the Galaxy Zoo project, which provides detailed visual morphological characterisations for Dark Energy Camera Legacy Survey images of galaxies within the SDSS DR8 footprint. Walmsley *et al.* [35] explore Transfer Learning that shows deep learning models trained to answer every Galaxy Zoo DECaLS question can learn representations of galaxies useful for new tasks. Walmsley *et al.* [35] have demonstrated success across three tasks. The first task was to identify similar morphology to a query galaxy using the free text tag assigned by humans (e.g. "#diffuse"). The second task identified the most "interesting" anomalies to a particular researcher with an accuracy of 100% using the most interesting anomalies identified in the Galaxy Zoo 2 data. The third task was to adapt a model to solve a new problem using only a limited number of newly-labelled galaxies.

B. THE HUBBLE TUNING FORK

The Hubble tuning fork characterisation scheme is an academically accepted method of classifying galaxies based on their physical structure. Furthermore, it is considered robust as it accounts for elliptical and both barred, and non-barred spiral galaxies [1], [31].

Outside of the Galaxy Zoo project, others have attempted to use machine learning for galaxy classification on other catalogues. Khalifa *et al.* [1] trained a CNN on a subset of the EFIGI catalogue [4], achieving an accuracy of 97.27% when classifying galaxies into the elliptical, spiral and irregular classes.

Hausen and Robertson [36], presented a deep learning framework, Morpheus, to classify images of galaxies. The training set comprised 7629 galaxies from the CANDELS survey in the GOODS South region. Hausen and Robertson [36] used four classes, namely disk, spheroid, irregular and point source/compact (for unresolved sources). The galaxy labels were a result of multiple experts voting on the characterisation of each galaxy. Morpheus was trained using all expert votes as training labels instead of other classification methods that used the correct labels and achieved an accuracy of Morpheus on the test set of 85.70%.

TABLE 1. Distribution of instances for Hubble classes in each catalogue.

Dataset	E	Irr	S0	SBa	SBb	SBc	Sa	Sb	Sc
EFIGI	.08	.11	.06	.06	.10	.06	.11	.18	.25
RSA	.14	.03	.14	.05	.08	.08	.12	.14	.22
GZ2-H	.40	.002	.08	.001	.08	.11	.002	.12	.20

C. TRANSFER LEARNING

Deep learning is an effective technique for making predictions, especially in image classification [37]. However, training deep learning models requires a significant amount of time, computational resources and data [37]. Transfer Learning is one solution to mitigate some of these limitations and has been shown in the literature to be highly effective in image classification problems [37]. Transfer learning involves using a deep learning model pre-trained on one dataset to initialise training on another dataset, with the two models most likely drawn from the same or a similar domain [37], [38].

D. THE REVISED SHAPLEY-AMES AND EFIGI CATALOGUES

The Revised Shapley-Ames catalogue contains 1 249 images of galaxies and their Hubble types [5]. Compiling a collection of the images from this catalogue proved challenging, as there are no readily available sources containing the subset of images. Therefore, we sourced all the images individually from the SDSS. In the experiments described in this paper, the RSA catalogue serves as an additional, expertly labelled test set.

The EFIGI catalogue, a subset of the *Third Reference Catalogue of Bright Galaxies* [4], contains 4488 images of galaxies. The catalogue uses the de Vaucouleurs system to classify galaxies, an extension of the original Hubble tuning fork expanded to include additional classes [4]; however, it is possible to map the expanded de Vaucouleurs system to the original Hubble tuning fork [4]. Since the EFIGI catalogue does not specify specific elliptical types, we grouped all ellipticals into type E, as shown in Table 1 when testing on the RSA catalogue [4]. We then used the EFIGI catalogue to obtain a dataset of galaxies with corresponding expertly labelled Hubble types. Subsequently, the datasets allow us to draw a comparison between the generalisation to the RSA catalogue of the GZ2-H crowd-sourced and the expert EFIGI catalogue.

Table 1 provides a consolidated view of the percentages of each of the Hubble classes shown in Figure 1 for each dataset used. For example, the table shows that the Irr, SBa and Sa classes are severely under-represented in the crowd-sourced

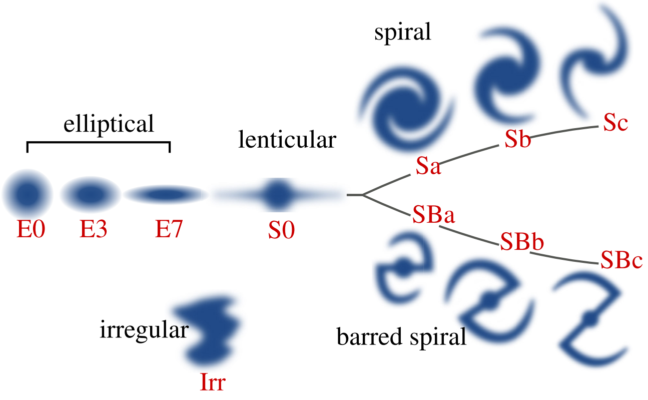


FIGURE 1. **Hubble tuning fork used in this research for the 9 and 11 classes.**

GZ2-H data, a limitation that is not present in the expert EFIGI and RSA catalogues. This under-representation shows the difficulty in classifying these galaxies for crowd-sourced human observers, evidenced by the class imbalance.

E. DEEP METRIC LEARNING

The following process summarises a machine learning classification problem: training data is fed into a network, following which class probabilities are outputted. This method requires large amounts of training samples similar to the ones expected when testing the model. On the other hand, metric learning proposes the idea of learning a complex, non-linear mapping [16]. This mapping maps high-dimensional input data to a lower-dimension manifold called an embedding space [16]. The mapping typically uses pairwise distances (such as the Euclidean norm or cosine similarity) as a metric [16]. Higher-dimension input data (e.g. images) trains a deep encoder. The deep encoder learns to map similar points (i.e. points with a low euclidean norm or cosine similarity) close to each other within the embedded space. The loss functions used to train the deep encoder are called pairwise losses, which penalise the model so that it encourages small distances for training samples within the same class and large distances for training samples with different classes [16].

Siamese Networks enable the implementation of Deep Metric Learning. A Siamese Network is an architecture with two or more parallel neural networks that share weights, and each network takes as input a different instance and compares the outputs to provide a measure of the similarity between the inputs [39].

In the case of image classification, these identical networks might be CNN's. Using samples of images during training, the model will learn embeddings or feature vectors that effectively extract latent features from the input images. When classifying a new image, the network can be used to produce the embeddings for all the training images and the new image. The distances between the query image embedding and the training image embeddings are then calculated. The query image will share the class of the training image closest to it (i.e. its nearest neighbour). When labelling a new image, the MAP@R value is calculated, which is a measure of how many of the top R nearest neighbours share the same class as the new image being classified.

The two sampling methods for training metric learning models are the triplet and pair mining functions [16].

- In triplet mining, each training sample, also known as the anchor (A), is fed into the network along with two other samples: A positive example (P), whose class/label will be the same as the anchors', and a negative example (N), whose class/label will be different to the anchors'.
- In pair mining, the network accepts samples in two pairs: The first pair contains a training sample, the anchor (A), and a positive sample (P), whose class/label is the same as the anchors'. The second pair will contain the same anchor A and a negative sample (N), whose class/label will differ from the anchors'.
- During training, the pairwise loss function penalises the network to decrease the distance (e.g. Euclidean norm) between the anchor and positive embedding spaces at each iteration (i.e. the embeddings for samples in the same class will move closer towards one another). In contrast, the distance between the anchor and the negative embedding spaces will increase at each iteration (i.e. the embeddings for samples in different classes will move further away from one another).

II. METHODOLOGY

The image pre-processing techniques applied to the Galaxy Zoo 2, EFIGI and RSA data were: cropping the image to 224 × 224 pixels, resizing the image to 75 × 75 pixels, automated augmentation techniques (such as random horizontal and vertical flips) and normalising the data. After performing image pre-processing, the data was split into train and validation subsets using a stratified approach (for the classification tasks). We cross-referenced the RSA galaxies with the EFIGI and GZ2-H sets and found that 420 galaxies were in both the RSA and EFIGI sets. To prevent training on galaxies from the RSA catalogue, we removed the overlapping samples from our EFIGI set before training, leaving us with 4068 images in the EFIGI catalogue.

A. GALAXY CLASSIFICATION: CROWD-SOURCED VS EXPERT LABELS

A ResNet50 model was trained to predict the 37-class response vectors in the Galaxy Zoo 2 data.The dataset was randomly split into train, test and validation subsets. The model used SGD during training with a learning rate of 0.1 and a momentum value of 0.9. The trained model is then used to predict the 37-class vectors in the Galaxy Zoo 2 data as initialisation for two models; one trained to predict the Hubble types of galaxies, provided in the Galaxy Zoo 2 data [8], and one trained to predict the Hubble types of galaxies in the EFIGI catalogue [4]. Both transfer learning models used the Adam optimisation algorithm during training with a learning rate of 1e^{-5} and a weight decay value of 5e^{-5}. The number of epochs was determined using early-stopping criteria on the respective validation losses. The model trained to predict the Galaxy Zoo 2 Hubble types trained for 178 epochs. The model trained to predict the EFIGI Hubble types trained for 104 epochs.

B. DEEP METRIC LEARNING

Boudiaf *et al.* [16], described a set of experiments to compare the effectiveness of the Cross-Entropy loss combined with Label Smoothing against the current state-of-the-art metric learning methods using benchmark classification problems. The benchmark problems included in the study are the Stanford Cars Dataset [40], the Caltech-UCSD Birds-200-2011 (CUB) Dataset [41], the Stanford Online Products [42] and the In-Shop Clothes Retrieval [43].

Following the work of Boudiaf *et al.* [16], a ResNet50 model was trained on both the EFIGI and GZ2-Hubble (GZ2-H) datasets using the Cross-Entropy loss combined with Label Smoothing. The one-hot encoded vectors were smoothed according to the same rules used by Boudiaf *et al.* [16]: For the positive class, i.e. 1, the class probability was set to:

$$1 - \epsilon, \tag{1}$$

and for the other classes, i.e. 0, the class probability was set to:

$$\frac{\epsilon}{K-1}, \tag{2}$$

where $\epsilon = 0.1$ and K represents the number of classes which, in our case, was 9 for the EFIGI catalogue and 11 for the GZ2-H dataset. These models are used to train metric learning models to investigate if using metric learning techniques for fine-tuning the Label Smoothing Cross-Entropy models would improve the accuracy and mean average precision (MAP).

The second-to-last layer of the ResNet50 has a size of 128, representing the size of the embedding space used to train the metric learning models. To determine the optimal size for the embedding space, various sizes were tested (512, 128 and 37), with 128 proving the most effective for the EFIGI catalogue and GZ2-H dataset. The size of the embedding space is kept at 128 in the subsequent metric learning models. Because the Galaxy Zoo 2 has 11 Hubble classes [8], and the EFIGI has 9 [4], the Galaxy Zoo 2 models are tested on both 9 and 11 Hubble classes for comparison.

Both triplet and pair mining were used for training models on the GZ2-H and EFIGI datasets and the same pairwise loss functions were used for both datasets. For the pairwise loss functions contrastive [44], triplet [45], Proxy-NCA [46], Proxy Anchor and Normalised-Softmax [16] losses are used. 5-fold cross-validation was used to reduce dataset dependence when training the models and the average MAP@R and accuracy values were calculated when testing the RSA catalogue.

TABLE 2. Datasets metadata.

Paper	Dataset	Samples size	# Hubble classes	Train/test/validation split (%)
[3]	GZ1	61578	-	70 / 15 / 15
[8]	GZ2-H	239573	9, 11	70 / 15 / 15
[4]	EFIGI cat.	4488	9	70 / 15 / 15
[5]	RSA cat.	1249	9, 11	Unseen test set

Cai *et al.* [2] described a deep learning framework, *DeepGalaxy*, trained to predict the timescales at which galaxies merge (i.e. the time it would take two galaxies to collide) based on their morphology. *DeepGalaxy* consists of a fully convolutional auto-encoder (FCAE) which generates activation maps at its 3-D latent-space. *DeepGalaxy* also consists of a variational autoencoder (VAE), which compresses the activation maps into a 1-D vector. Finally, *DeepGalaxy* consists of a classifier that generates labels from the activation maps [2]. The dataset was made up of 35784 images generated from 36 N-body simulations [2], and is available online. We followed their methodology and train a ResNet50 model on the simulated dataset. This model was used to generate the 1-D feature vectors for the EFIGI catalogue. A multi-layer perceptron (MLP) which used these 1-D vectors input and the one-hot encoded Hubble labels of the EFIGI as the output (targets) was trained. We then investigated the accuracy obtained by the MLP on both the unseen EFIGI and additional RSA test sets to compare with the accuracy obtained using our transfer learning model [11] and present the findings in Section IV.

Table 2 summarises the information on the datasets used for training and testing the models. All the models described were trained in Python 3.7 using PyTorch version 1.2.0 on an Intel® Core$^{\text{TM}}$ i7-7700 CPU @ 3.60 GHz, 16GB DDR4 2400 MHz and an Nvidia® GeForce® GTX 1060 6GB GDDR5 GPU. The utilised code for all experiments can be found on GitHub [47].

III. RESULTS

Table 4 summarises the datasets used and the results obtained for training and testing of the models in the metric learning experiments. A summary of the classification models' results and the information retrieval precision using the Label-Smoothing Cross-Entropy loss can be found in Table 3.

A. GALAXY CLASSIFICATION: CROWD-SOURCED VS. EXPERT LABELS

Table 3 includes the retrieval precision's (MAP@1 and MAP@R) for each of the considered classification techniques alongside the accuracy and F1 score. The models are trained using LS-CE and a Res50 architecture. The results are reported for the completely unseen RSA catalogue as a test set. As discussed in Section II-A the model trained on the GZ2-H dataset achieved an accuracy of 25.70% and an unweighted F_1 score of 0.2035 when tested on the RSA

TABLE 3. **Classification accuracy [11] and retrieval precision on RSA.**

Model init' on	Fine tuning	# Hubble classes	RSA accuracy	RSA F1	RSA MAP@1	RSA MAP@R[†]
GZ2	EFIGI	*9*	*29.52*	**0.2822**	**24.86**	**0.3119**
GZ2	GZ2-H	*9*	**30.50**	0.2383	*24.32*	*0.3101*
GZ2	GZ2-H	*11*	*25.70*	0.2035	*20.85*	*0.2704*

[†] The MAP@R value is calculated by taking the Average Precision for all R neighbours of a query, and then calculating the mean of that value.

catalogue using 11 Hubble classes [11]. Using 9 Hubble classes, this model achieved an accuracy of 30.50% and an unweighted F_1 score of 0.2383. The same model trained on the EFIGI catalogue achieved an accuracy of 29.52% and an unweighted F_1 score of 0.2822 when tested on the RSA catalogue using 9 Hubble classes. The results show that the best accuracy is obtained when fine-tuning on GZ2-H with nine classes. However, the best F1 and retrieval are for fine-tuning on EFIGI. These results demonstrate that better accuracy does not always lead to better retrieval and motivate the necessity for deep metric learning over classification and the further experimentation in Section III-B.

B. DEEP METRIC LEARNING

The results in this Section are for the metric learning models trained for similarity learning on the EFIGI and GZ2-H datasets and tested on the unseen RSA catalogue. 5-fold cross-validation was used to reduce dataset noise when training the models and for validation. Training the models used 4 folds, validation used the 5th and testing was done on the RSA catalogue. For each metric learning model, both pair and triplet mining are evaluated, and only the optimal mining function is presented. Table 4 provides a consolidated view of the retrieval results obtained in the metric learning experiments alongside the classification results.

Multiple models were trained on the GZ2-H dataset using the aforementioned pairwise loss functions for testing on the RSA 11-class data. The results in Table 4 show that the Normalised-Softmax loss function yielded the best accuracy and MAP@R values at 25.04% and 0.3102, respectively. However, these results are not the best results found overall.

When considering the RSA 9-class data, multiple models were trained and fine-tuned on the EFIGI and GZ2-H catalogue using the selected pairwise loss functions. The models were tested on the unseen RSA catalogue. The results in Table 4 show that fine-tuning on EFIGI gives better MAP@R values for all methods. Finally, Proxy-NCA loss fine-tuned on EFIGI has good accuracy and MAP@R at 29.03% and 0.3917, respectively. For fine-tuned on the GZ2-H catalogue, the Normalised-Softmax loss function gave the best accuracy at 30.88% with, Proxy-NCA having the highest MAP@R at 0.3577. Finally, it is worth noting that Proxy-NCA and Normalised-Softmax are the best two performing methods when comparing across all the experiments.

The presented results also include additional confusion matrices. These results are MAP@1 for the two best loss functions, Proxy-NCA and Normalised-Softmax, on the RSA 9-class dataset. Here MAP@1 is considered equivalent to the accuracy of each class. The results also include the worst performing of these two loss functions for the GZ2-H 11-class dataset. The GZ2-H 11-class confusion matrix is included for further dissection of the E class concerning the differences to the 9-class case.

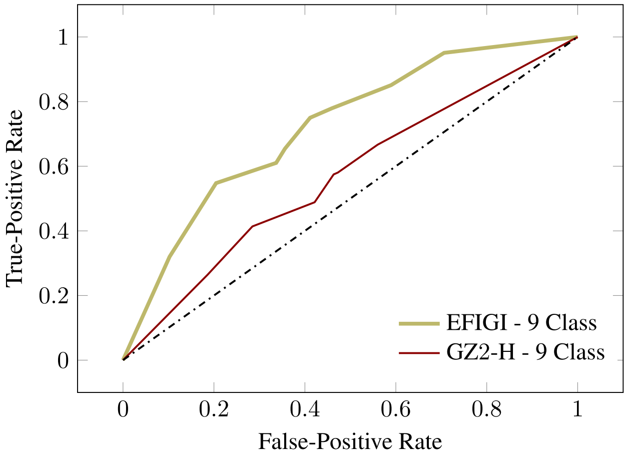


FIGURE 2. **Receiver operating characteristic curves for the EFIGI 9-class Triplet/Normalised-Softmax model and GZ2-H 9-class Triplet/Normalised-Softmax model.**

In summary, the confusion matrices included are as follows. The results for actual vs predicted on the RSA 9-class labels using the Proxy-NCA loss on the GZ2-H dataset are in Table 5 and for the EFIGI dataset in Table 7. The same results for Normalised-Softmax on GZ2-H are in Table 6 and for EFIGI in Table 8. Finally, Table 9 shows the MAP@1 actual vs predicted labels for the GZ2-H 11 class using the Proxy-NCA model.

The final results are presented in Figure 2 that shows a Receiver Operating Characteristic curve (ROC) for the two best models for each fine tuning scenario. We use this result to further analyse the true-positive vs. false-positive rates of these methods.

IV. DISCUSSION

1) UNIFYING AND CONTRASTING CLASSIFICATION AND METRIC LEARNING

One difficulty with current research is that comparison across methods is difficult due to the varying number of classes used in different studies and available in the different datasets. Further, most existing results have considered classification quality rather than the quality of the embeddings achieved as measured by retrieval precision.

Variawa *et al.* [11] achieve a classification accuracy of 85.22% compared to the 65.20% obtained by Barchi *et al.* [31] on the Galaxy Zoo 2 Hubble (GZ2-H) unseen test set. They also achieve an accuracy of 97.76% compared to the 97.27% obtained by Khalifa *et al.* [1] on the EFIGI unseen test set. Variawa *et al.* [11] evaluated these fine-tuning classification task on the unseen 9 and 11 classes RSA catalogue

TABLE 4. Deep metric learning retrieval results showing the datasets, loss and mining functions, accuracy and MAP@R with embedding size 128.

Model init' on	Fine tuning	Number classes	Loss type	Loss function	Mining function	RSA MAP@1	RSA MAP@R†
GZ2	EFIGI	9	Classification	LS-CE	-	24.86	0.3119
EFIGI (LS-CE)	EFIGI	9	Metric Learning	Contrastive	Pair	27.97	0.3859
EFIGI (LS-CE)	EFIGI	9	Metric Learning	Norm-S'max	Triplet	**30.05**	0.3889
EFIGI (LS-CE)	EFIGI	9	Metric Learning	P-Anchor	Triplet	**29.32**	**0.3953**
EFIGI (LS-CE)	EFIGI	9	Metric Learning	P-NCA	Pair	29.03	**0.3917**
EFIGI (LS-CE)	EFIGI	9	Metric Learning	Triplet	Triplet	28.98	0.3898
GZ2	GZ2-H	9	Classification	LS-CE	-	24.32	0.3101
GZ2-H (LS-CE)	GZ2-H	9	Metric Learning	Contrastive	Pair	26.84	0.3362
GZ2-H (LS-CE)	GZ2-H	9	Metric Learning	Norm-S'max	Triplet	**30.88**	**0.3483**
GZ2-H (LS-CE)	GZ2-H	9	Metric Learning	P-Anchor	Triplet	29.48	0.3350
GZ2-H (LS-CE)	GZ2-H	9	Metric Learning	P-NCA	Pair	**30.33**	**0.3577**
GZ2-H (LS-CE)	GZ2-H	9	Metric Learning	Triplet	Triplet	26.42	0.3328
GZ2	GZ2-H	11	Classification	LS-CE	-	20.85	0.2704
GZ2-H (LS-CE)	GZ2-H	11	Metric Learning	Contrastive	Pair	23.86	0.3032
GZ2-H (LS-CE)	GZ2-H	11	Metric Learning	Norm-S'max	Pair	**25.04**	**0.3102**
GZ2-H (LS-CE)	GZ2-H	11	Metric Learning	P-Anchor	Pair	24.72	0.2909
GZ2-H (LS-CE)	GZ2-H	11	Metric Learning	P-NCA	Triplet	**24.45**	**0.2955**
GZ2-H (LS-CE)	GZ2-H	11	Metric Learning	Triplet	Triplet	22.50	0.2900

† The MAP@R value is calculated by taking the Average Precision for the top R neighbours for each sample, and then calculating the mean of that value.

TABLE 5. Revised Shapley-Ames actual hubble type vs predicted hubble type - Galaxy Zoo 2 (9 Classes) Pair/Proxy-NCA model - MAP@1.

		Predicted Type								
		E	Irr	S0	SBa	SBb	SBc	Sa	Sb	Sc
Actual Type	E	.24	.0	.07	.0	.05	.0	.0	.27	.37
	Irr	.32	.0	.02	.0	.07	.2	.0	.07	.32
	S0	.21	.0	.14	.0	.07	.03	.0	.24	.3
	SBa	.32	.0	.03	.0	.03	.06	.0	.35	.2
	SBb	.42	.0	.05	.0	.09	.03	.0	.09	.31
	SBc	.33	.0	.07	.0	.13	.07	.0	.08	.33
	Sa	.27	.0	.08	.01	.08	.14	.0	.18	.24
	Sb	.24	.0	.03	.0	.05	.25	.0	.09	.34
	Sc	.24	.0	.02	.0	.08	.27	.0	.05	.33

TABLE 6. Revised Shapley-Ames actual hubble type vs predicted hubble type - Galaxy Zoo 2 (9 Classes) Triplet/Normalised-Softmax model - MAP@1.

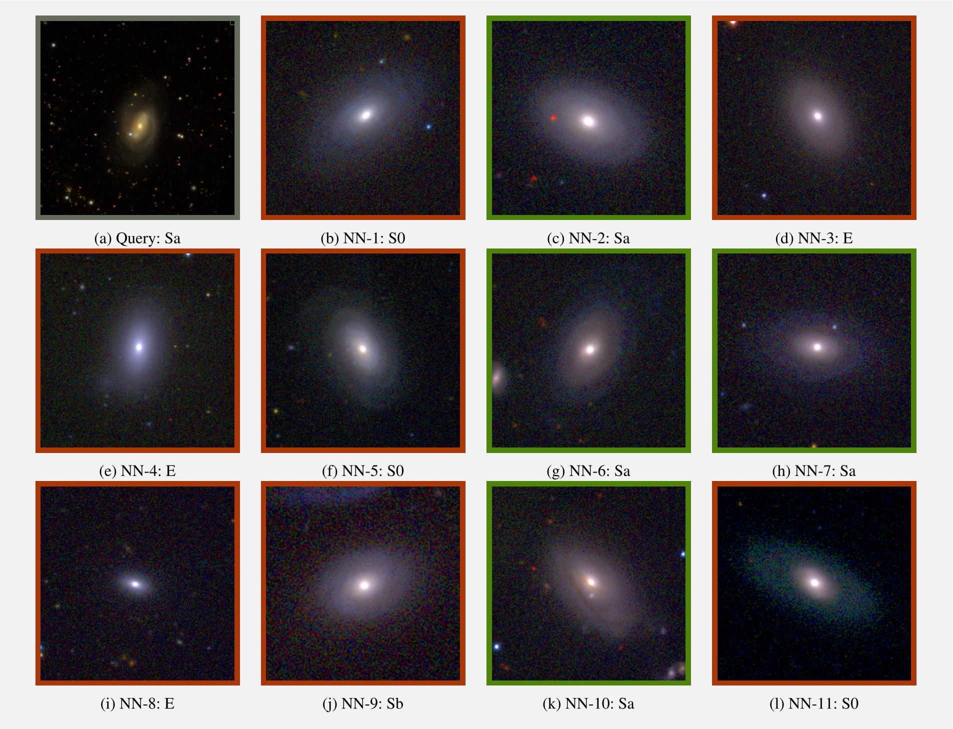
		Predicted Type								
		E	Irr	S0	SBa	SBb	SBc	Sa	Sb	Sc
Actual Type	E	.38	.0	.07	.01	.03	.09	.0	.17	.25
	Irr	.41	.0	.06	.01	.03	.07	.0	.14	.27
	S0	.44	.0	.06	.02	.03	.08	.0	.11	.25
	SBa	.48	.0	.05	.01	.07	.08	.0	.11	.2
	SBb	.45	.0	.06	.0	.06	.07	.0	.11	.24
	SBc	.39	.0	.08	.0	.05	.06	.0	.17	.25
	Sa	.47	.0	.05	.01	.05	.08	.0	.1	.24
	Sb	.46	.0	.04	.0	.05	.06	.0	.11	.26
	Sc	.4	.0	.06	.0	.06	.06	.0	.17	.25

and achieve the replicated accuracy's presented in Table 3. As such, Variawa *et al.* [11] have achieved either a significant improvement or parity in the aforementioned classification tasks, allowing this study to consider their approach as a baseline.

When considering the results in Table 3 we note that the model initialised on GZ2 and fine-tuned on GZ2-H does best in terms of classification accuracy. However, considering the true retrieval performance (MAP@R), we note that fine-tuning EFIGI is slightly improved over fine-tuning on GZ2-H. This result supports further investigation into metric learning for improved embeddings for retrieval tasks in the following sub-section.

2) CROWD-SOURCED VS EXPERT LABELS

For the results described in Section III-A, the MAP@1 and MAP@R values are provided in Table 4. The values provided aim to contrast previous research with the metric learning experiments described in Section II-B. Previously, Boudiaf *et al.* [16], show that Label Smoothing Cross-Entropy loss serves as a good weight initialisation for metric learning methods. Further, these results demonstrate that training a ResNet50 using Label Smoothing Cross-Entropy and then fine-tuning the model using metric learning techniques improved on previous results, described in Section III-A. An example image from the RSA catalogue, with Hubble type Sa, is shown in Figure 3 and the corresponding 11 nearest neighbours in the EFIGI catalogue. The model further demonstrates state-of-the-art classification and metric learning of a galaxy's Hubble type for both the crowd-sourced and expert datasets.


(a) Query: Sa

(b) NN-1: S0

(c) NN-2: Sa

(d) NN-3: E

(e) NN-4: E

(f) NN-5: S0

(g) NN-6: Sa

(h) NN-7: Sa

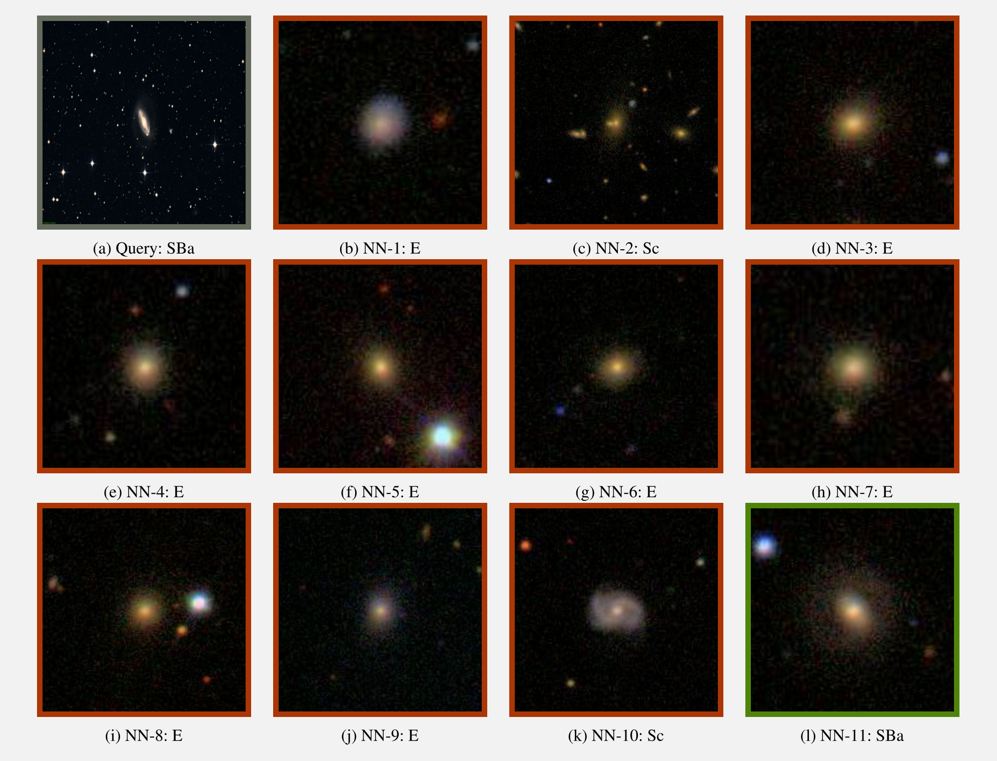
(i) NN-8: E

(j) NN-9: Sb

(k) NN-10: Sa

(l) NN-11: S0

FIGURE 3. An example image from the RSA catalogue, Hubble type Sa, with the top 11 nearest neighbours (NN) from the EFIGI Triplet/Proxy-NCA model.


(a) Query: SBa

(b) NN-1: E

(c) NN-2: Sc

(d) NN-3: E

(e) NN-4: E

(f) NN-5: E

(g) NN-6: E

(h) NN-7: E

(i) NN-8: E

(j) NN-9: E

(k) NN-10: Sc

(l) NN-11: SBa

FIGURE 4. An example image from the RSA catalogue, Hubble type SBa, with the top 11 nearest neighbours (NN) from the GZ2-H Pair/Normalised-Softmax model.

The *DeepGalaxy* framework described by Cai *et al.* [2] was shown to be state-of-the-art at predicting the timescales at which galaxies would merge (i.e. the time it would take two galaxies to collide). However, it also demonstrated that

TABLE 7. Revised Shapley-Ames actual hubble type vs predicted hubble type - EFIGI (9 Classes) Triplet/Proxy-NCA model - MAP@1.

Actual Type \ Predicted Type	E	Irr	S0	SBa	SBb	SBc	Sa	Sb	Sc
E	.71	.05	.17	.02	.0	.0	.05	.01	.0
Irr	.37	.24	.1	.0	.0	.07	.02	.07	.12
S0	.47	.04	.2	.07	.0	.0	.15	.06	.01
SBa	.25	.06	.12	.15	.06	.02	.29	.02	.03
SBb	.2	.07	.08	.09	.19	.02	.14	.12	.08
SBc	.22	.11	.05	.03	.11	.13	.06	.1	.2
Sa	.27	.05	.21	.07	.01	.0	.28	.07	.04
Sb	.2	.09	.07	.08	.08	.0	.17	.21	.1
Sc	.08	.07	.05	.03	.09	.05	.08	.24	.3

TABLE 8. Revised Shapley-Ames actual hubble type vs predicted hubble type - EFIGI (9 Classes) Triplet/Normalised-Softmax model - MAP@1.

Actual Type \ Predicted Type	E	Irr	S0	SBa	SBb	SBc	Sa	Sb	Sc
E	.7	.07	.14	.01	.0	.0	.06	.01	.01
Irr	.24	.28	.07	.0	.02	.08	.05	.06	.19
S0	.43	.09	.2	.04	.0	.01	.15	.05	.03
SBa	.24	.1	.18	.17	.06	.02	.15	.05	.04
SBb	.15	.1	.07	.09	.16	.04	.14	.14	.11
SBc	.14	.24	.03	.03	.1	.1	.06	.1	.2
Sa	.29	.09	.16	.07	.03	.01	.21	.11	.04
Sb	.14	.13	.09	.05	.07	.02	.14	.21	.15
Sc	.1	.1	.04	.03	.05	.05	.08	.21	.34

TABLE 9. Revised Shapley-Ames actual hubble type vs predicted hubble type - Galaxy Zoo 2 (11 Classes) Triplet/Proxy-NCA model - MAP@1.

Actual Type \ Predicted Type	E0	E3-5	E7	Irr	S0	SBa	SBb	SBc	Sa	Sb	Sc
E0	.19	.05	.15	.0	.01	.0	.05	.06	.0	.18	.31
E3-5	.17	.04	.26	.0	.0	.0	.03	.11	.0	.2	.19
E7	.21	.12	.04	.0	.0	.0	.0	.12	.0	.21	.29
Irr	.39	.1	.1	.0	.0	.0	.07	.24	.0	.07	.02
S0	.18	.05	.08	.0	.01	.0	.02	.17	.0	.2	.29
SBa	.09	.11	.06	.0	.12	.0	.06	.11	.0	.17	.28
SBb	.16	.27	.02	.0	.02	.0	.05	.19	.0	.03	.26
SBc	.2	.2	.01	.0	.01	.0	.12	.21	.0	.1	.16
Sa	.16	.04	.07	.0	.03	.0	.07	.18	.0	.16	.29
Sb	.18	.17	.01	.0	.0	.0	.06	.16	.0	.16	.26
Sc	.26	.13	.02	.0	.0	.0	.07	.15	.0	.21	.17

the *DeepGalaxy* framework did not transfer well to the task of classifying the Hubble types of galaxies. Specifically, on the unseen EFIGI test set, the model achieved an accuracy of 28.80% compared to the 85.22% reported by Variawa *et al.* [11]. On the RSA test set, the accuracy achieved using *DeepGalaxy* framework was 18.33% compared to the 29.52% reported by Variawa *et al.* [11].

The results in Table 4 show that the Proxy-NCA loss and Normalised-Softmax functions have marginally outperformed the other pairwise loss functions on both the 9 class and 11 datasets when retrieving Hubble types in the RSA catalogue:

- The best overall GZ2-H 9-class model trained using the Label Smoothing Cross-Entropy loss function. The triplet mining and Normalised-Softmax loss function fine-tuned the model to yield an accuracy of 30.88% and a MAP value of 0.3483.
- The second-best overall GZ2-H 9-class model trained using the Label Smoothing Cross-Entropy loss function. The pair mining and Proxy-NCA loss function fine-tuned the model to yield an accuracy of 30.33% and a MAP value of 0.3577.
- The best overall EFIGI 9-class model trained using the Label Smoothing Cross-Entropy loss function and then fine-tuned using the triplet mining function and Normalised-Softmax loss function, yielding an accuracy of 30.05% and a MAP value of 0.3889.
- The second-best overall EFIGI 9-class model trained using the Label Smoothing Cross-Entropy loss function and then fine-tuned using the pair mining function and Proxy-NCA loss function, yielding an accuracy of 29.03% and a MAP value of 0.3917.

The results show that metric learning has provided an improvement over the classification results presented in Table 3. Specifically, the best GZ2-H 9-class model achieved an accuracy of 30.88%, an improvement of 0.38%, and a MAP value of 0.3483, an increase of 0.0382. The best EFIGI 9-class model produced an accuracy of 30.05%, an improvement of 0.53%, and a MAP value of 0.3889, an increase of 0.077. These improvements highlight that metric learning was better suited, than the classification approach, to deal with the class imbalance of the GZ2-H dataset, mentioned in Section I-D.

Finally, through extensive experimentation, the results have shown that crowd-sourced data is less effective for galaxy retrieval and embeddings in metric learning. For example, the models trained on the crowd-sourced GZ2-H dataset produced MAP@R values much lower than the models trained on the expertly labelled EFIGI catalogue, as shown in Table 4. These results support the findings by Variawa *et al.* [11] who showed the same was true for classification.

Turning to the confusion matrices allow us to further investigate the under-performance cause when using the crowd-sourced GZ2-H dataset. Studying the confusion matrices in Table 5 and 6 shows that the models trained on the GZ2-H dataset were not able to predict the Irr, SBa and Sa classes. These models were unable able to predict these classes due to these 3 classes being severely under-represented in the GZ2-H dataset, as shown in Table 1, relative to the other Hubble types.

Further investigation showed that galaxies with an actual label Sa were often classified as galaxy types Sb and Sc (the MAP@1 values for Sb and Sc were the second and third highest at 0.18 and 0.24, respectively), as seen in Table 5 and 6. The misclassification of the Hubble type Sa as Sb and Sc indicates that the Sb and Sc classes follow the Sa class on the Hubble tuning fork as seen in Figure 1. Figure 4 show an example image from the RSA catalogue, with Hubble type SBa (another under-represented class in the GZ2-H dataset) and the nearest neighbours from the GZ2-H dataset. Figure 4 illustrates that the model, trained using the pair mining function with the Normalised-Softmax loss function, cannot easily pick out sample images with similar structures to the query. Table 7 and 8 shows that this problem is not prevalent in the models trained on the EFIGI catalogue. Since the models trained on the EFIGI catalogue did not have the same problem, it can be concluded that using an expertly labelled catalogue of galaxies is better for automating galaxy classification using machine learning than using a crowd-sourced dataset.

Studying the ROC curves in Figure 2 is another indication (substantiating the claim above) that an expertly labelled catalogue of galaxies is better than crowd-sourced data for automating galaxy classification using machine learning. As seen in Figure 2, the ROC curve for the EFIGI 9-class rises above the GZ2-H 9-class curves. These ROC curves illustrate that the Triplet/Normalised-Softmax model trained on the EFIGI catalogue was better at predicting the Hubble types of galaxies in the RSA set than the Triplet/Normalised-Softmax model trained on the GZ2-H 9-class data.

Finally, studying the confusion matrices in Table 9 shows that the models trained on the GZ2-H dataset were not able to predict the Irr, SBa and Sa classes but were able to disambiguate E0, E3-5 and E7. This result further reinforces that the models trained on the GZ2-H dataset were not able to predict the Irr, SBa and Sa classes due to these 3 classes being severely under-represented in the GZ2-H dataset, as shown in Table 1.

V. CONCLUSION

The type and formation of galaxies often offer clues and insights into the origin and evolution of the universe, making galaxy classification a key area of study. As a result of the number of images of galaxies available, training machine learning models on crowd-sourced data is an increasingly utilised approach to ''automate'' classification.

Given the popularity of crowd-sourced labels for data, it is necessary to investigate how well machine learning models trained on this data can generalise to the expertly labelled unseen data. An example of expertly labelled unseen data is our RSA catalogue labelled with the Hubble tuning fork system.

Confusion matrices and ROC curves have shown that galaxy classification can be better automated by training deep learning models on the expertly labelled EFIGI catalogue instead of the crowd-sourced GZ2-H dataset. It has also been shown that a model trained using Label Smoothing Cross-Entropy can be fine-tuned using metric learning techniques to outperform the state-of-the-art in galaxy retrieval. This result, however, is in part due to the models trained on the GZ2-H being unable to predict the Irr, SBa and Sa classes, as shown in Tables 9 and 5, because these classes are severely underrepresented in the GZ2-H data, as shown in Table 1. Table 1 shows that in the EFIGI and RSA data, these classes are not underrepresented.

These results lead us to three significant findings. Firstly, caution should be exercise when using crowd-sourced labels, especially for difficult minority classes. Secondly, the use of similarity or deep metric learning is a viable approach to improving the state-of-the-art in galaxy classification and related tasks. Thirdly, transfer learning from crowd-sourced labelled data to expert-labelled data leads to significant improvements accuracy.

Finally, although the results bring into question the efficacy of using crowd-sourced labels alone, they do not preclude the importance of further investigating how semi-supervised deep learning and cross/transfer learning might exploit all the data available.

REFERENCES

[1] N. Eldeen M. Khalifa, M. Hamed N. Taha, A. E. Hassanien, and I. M. Selim, ''Deep galaxy: Classification of galaxies based on deep convolutional neural networks,'' 2017, *arXiv:1709.02245*.

[2] M. X. Cai, J. Bédorf, V. A. Saletore, V. Codreanu, D. Podareanu, A. Chaibi, and P. X. Qian, ''DeepGalaxy: Deducing the properties of galaxy mergers from images using deep neural networks,'' 2020, *arXiv:2010.11630*.

[3] C. Lintott, K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, E. Edmondson, K. Masters, R. C. Nichol, M. J. Raddick, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, ''Galaxy zoo 1: Data release of morphological classifications for nearly 900 000 galaxies: Galaxy zoo,'' *Monthly Notices Roy. Astronomical Soc.*, vol. 410, no. 1, pp. 166–178, Jan. 2011.

[4] G. De Vaucouleurs, *3rd Reference Catalogue Bright Galaxies*, vol. 3. New York, NY, USA: Springer, 2013.

[5] A. Sandage and G. Tammann, *A Revised Shapley-Ames Catalog of Bright Galaxies*. Washington, DC, USA: Carnegie Inst. Washington, 1981.

[6] M. Jiménez, M. T. Torres, R. John, and I. Triguero, ''Galaxy image classification based on citizen science data: A comparative study,'' *IEEE Access*, vol. 8, pp. 47232–47246, 2020.

[7] 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, 2019.

[8] R. E. Hart, S. P. Bamford, K. W. Willett, K. L. Masters, C. Cardamone, C. J. Lintott, R. J. Mackay, R. C. Nichol, C. K. Rosslowe, B. D. Simmons, and R. J. Smethurst, ''Galaxy zoo: Comparing the demographics of spiral arm number and a new method for correcting redshift bias,'' *Monthly Notices Roy. Astronomical Soc.*, vol. 461, no. 4, pp. 3663–3682, Oct. 2016.

[9] M. Franzen, L. Kloetzer, M. Ponti, J. Trojan, and J. Vicens, ''Machine learning in citizen science: Promises and implications,'' *Sci. Citizen Sci.*, vol. 4, p. 183, Oct. 2021.

[10] D. P. Sullivan, C. F. Winsnes, L. Åkesson, M. Hjelmare, M. Wiking, R. Schutten, L. Campbell, H. Leifsson, S. Rhodes, A. Nordgren, K. Smith, B. Revaz, B. Finnbogason, A. Szantner, and E. Lundberg, ''Deep learning is combined with massive-scale citizen science to improve large-scale image classification,'' *Nature Biotechnol.*, vol. 36, no. 9, pp. 820–828, Oct. 2018.

[11] M. Z. Variawa, T. van Zyl, and M. Woolway, ''Comparing generalisation using crowd-sourced vs expert labels for galaxies classification,'' in *Proc. ISCMI*, 2020, pp. 158–162.

[12] X. Li, L. Yu, C.-W. Fu, M. Fang, and P.-A. Heng, ''Revisiting metric learning for few-shot image classification,'' *Neurocomputing*, vol. 406, pp. 49–58, May 2020.

[13] Z. Ding and Y. Fu, ‘‘Robust transfer metric learning for image classification,’’ *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 660–670, Feb. 2017.

[14] X. Zhe, S. Chen, and H. Yan, ‘‘Directional statistics-based deep metric learning for image classification and retrieval,’’ *Pattern Recognit.*, vol. 93, pp. 113–123, Sep. 2019.

[15] H. Wang, L. Feng, J. Zhang, and Y. Liu, ‘‘Semantic discriminative metric learning for image similarity measurement,’’ *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1579–1589, Aug. 2016.

[16] M. Boudiaf, J. Rony, I. M. Ziko, E. Granger, M. Pedersoli, P. Piantanida, and I. B. Ayed, ‘‘A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses,’’ in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 548–564.

[17] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, ‘‘Ranked list loss for deep metric learning,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5207–5216.

[18] F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff, ‘‘Deep metric learning to rank,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1861–1870.

[19] T. L. Van Zyl, M. Woolway, and B. Engelbrecht, ‘‘Unique animal identification using deep transfer learning for data fusion in Siamese networks,’’ in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Jul. 2020, pp. 1–6.

[20] N. Dlamini and T. L. van Zyl, ‘‘Author identification from handwritten characters using Siamese CNN,’’ in *Proc. Int. Multidisciplinary Inf. Technol. Eng. Conf. (IMITEC)*, Nov. 2019, pp. 1–6.

[21] K. Ohri and M. Kumar, ‘‘Review on self-supervised image recognition using deep neural networks,’’ *Knowl.-Based Syst.*, vol. 224, Apr. 2021, Art. no. 107090.

[22] O. Ciga, T. Xu, and A. L. Martel, ‘‘Self supervised contrastive learning for digital histopathology,’’ *Mach. Learn. Appl.*, vol. 7, Mar. 2022, Art. no. 100198.

[23] Y. P. Sun and M. Zhang, ‘‘Compositional metric learning for multi-label classification,’’ *Frontiers Comput. Sci.*, vol. 15, Dec. 2020, Art. no. 155320.

[24] M. A. Hayat, G. Stein, P. Harrington, Z. Luki, and M. Mustafa, ‘‘Self-supervised representation learning for astronomical images,’’ *Astrophys. J. Lett.*, vol. 911, no. 2, p. L33, Apr. 2021.

[25] P. Jia, R. Ning, R. Sun, X. Yang, and D. Cai, ‘‘Data-driven image restoration with option-driven learning for big and small astronomical image data sets,’’ *Monthly Notices Roy. Astronomical Soc.*, vol. 501, no. 1, pp. 291–301, Feb. 2021.

[26] H. Shao and D. Zhong, ‘‘Towards open-set touchless palmprint recognition via weight-based meta metric learning,’’ *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108247.

[27] 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, Jun. 2015.

[28] V. Lukic and M. Bruggen, ‘‘Galaxy classification with deep learning,’’ in *Proc. Int. Astronomical Union*, 2016, pp. 217–220.

[29] R. Katebi, Y. Zhou, R. Chornock, and R. Bunescu, ‘‘Galaxy morphology prediction using capsule networks,’’ 2018, *arXiv:1809.08377*.

[30] M. Z. Variawa, T. L. van Zyl, and M. Woolway, ‘‘A rules-based and transfer learning approach for deriving the Hubble type of a galaxy from the galaxy zoo data,’’ in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Jul. 2020, pp. 1–7.

[31] P. Barchi, R. de Carvalho, R. Rosa, R. Sautter, M. Soares-Santos, B. Marques, E. Clua, T. Gonçalves, C. de Sá-Freitas, and T. Moura, ‘‘Machine and deep learning applied to galaxy morphology-a comparative study,’’ *Astron. Comput.*, vol. 30, Jan. 2020, Art. no. 100334.

[32] R. Gupta, P. Srijith, and S. Desai, ‘‘Galaxy morphology classification using neural ordinary differential equations,’’ *Astron. Comput.*, vol. 38, Jan. 2022, Art. no. 100543.

[33] M. Reza, ‘‘Galaxy morphology classification using automated machine learning,’’ *Astron. Comput.*, vol. 37, Oct. 2021, Art. no. 100492.

[34] M. Walmsley, C. Lintott, T. Geron, S. Kruk, C. Krawczyk, K. W. Willett, S. Bamford, L. S. Kelvin, L. Fortson, Y. Gal, W. Keel, K. L. Masters, V. Mehta, B. D. Simmons, R. Smethurst, L. Smith, E. M. Baeten, and C. Macmillan, ‘‘Galaxy zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies,’’ 2021, *arXiv:2102.08414*.

[35] M. Walmsley, A. M. M. Scaife, C. Lintott, M. Lochner, V. Etsebeth, T. Géron, H. Dickinson, L. Fortson, S. Kruk, K. L. Masters, K. Bharadwaj Mantha, and B. D. Simmons, ‘‘Practical galaxy morphology tools from deep supervised representation learning,’’ 2021, *arXiv:2110.12735*.

[36] R. Hausen and B. Robertson, ‘‘Morpheus: A deep learning framework for pixel-level analysis of astronomical image data,’’ 2019, *arXiv:1906.11248*.

[37] K. Weiss, T. M. Khoshgoftaar, and D. Wang, ‘‘A survey of transfer learning,’’ *J. Big Data*, vol. 3, no. 1, p. 9, Dec. 2016.

[38] S. J. Pan and Q. Yang, ‘‘A survey on transfer learning,’’ *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[39] R. Hadsell, S. Chopra, and Y. LeCun, ‘‘Dimensionality reduction by learning an invariant mapping,’’ in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1735–1742.

[40] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, ‘‘3D object representations for fine-grained categorization,’’ in *Proc. 4th IEEE Workshop 3D Represent. Recognit.*, Sydney, NSW, Australia, Dec. 2004, pp. 554–561.

[41] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, ‘‘The CALTECH-UCSD birds-200-2011 dataset,’’ California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[42] O. Can, ‘‘Deep metric learning with alternating projections onto feasible sets,’’ 2019, *arXiv:1907.07585*.

[43] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, ‘‘DeepFashion: Powering robust clothes recognition and retrieval with rich annotations,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.

[44] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, ‘‘Multi-similarity loss with general pair weighting for deep metric learning,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5022–5030.

[45] G. Weifeng, ‘‘Deep metric learning with hierarchical triplet loss,’’ in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 269–285.

[46] J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov, ‘‘Neighbourhood components analysis,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1–8.

[47] M. Z. Variawa, T. van Zyl, and M. Woolway. (2020). *Galaxy Classification*. [Online]. Available: https://github.com/mohamedzayyan/galaxyclassification

MOHAMED ZAYYAN VARIAWA received the M.Sc. degree from the University of the Witwatersrand, Johannesburg, South Africa, in 2021. He is currently working as a Data Scientist, with his tasks focusing mainly on the development and productionization of machine learning models.

TERENCE L. VAN ZYL (Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science for his thesis on agent-based complex adaptive systems from the University of Johannesburg, South Africa. He holds the Nedbank Research and Innovation Chair of the University of Johannesburg, where he is currently a Professor with the Institute for Intelligent Systems. He is also an NRF-rated scientist and he has more than 15 years of experience researching and innovating large scale streaming analytics systems for government and industry. His research interests include data-driven science and engineering, prescriptive analytics, machine learning, meta-heuristic optimization, complex adaptive systems, high-performance computing, and artificial intelligence.

MATTHEW WOOLWAY received the Ph.D. degree in process engineering from the University of the Witwatersrand, Johannesburg, South Africa. He is currently an Industry Data Scientist and a Research Associate with the Faculty of Engineering and the Built Environment, University of Johannesburg. His research interests include computational intelligence, artificial intelligence, and optimization.