DISSERTATION

Galaxy classification using machine learning

Author:
Mohamed Zayyan VARIAWA
(852648)

Supervisors:
Prof. Terence VAN ZYL
Dr. Matthew WOOLWAY

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

submitted to
the Faculty of Science, in fulfilment of the requirements for the degree of
MSc in Computer Science

in the

School of Computer Science and Applied Mathematics

September 5, 2021

Declaration of Authorship

I, Mohamed Zayyan VARIAWA (852648), declare that this dissertation titled, "Galaxy classification using machine learning" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work.
- I have acknowledged all main sources of help.
- Where the dissertation is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: M Z Variawa

Date: September 5, 2021

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Abstract

Faculty of Science

School of Computer Science and Applied Mathematics

MSc in Computer Science

Galaxy classification using machine learning

by Mohamed Zayyan VARIAWA (852648)

An important area of study is galaxy classification, as the type and formation of galaxies often offer insights into the origin and evolution of the universe. The majority of classifications comes from human experts manually inspecting and labelling images of galaxies. Owing to the increased availability of images of galaxies, researchers coupled machine learning with crowd-sourced labels to automate the process of galaxy classification to save time spent by astronomers performing manual classification. However, studying the generalisation of these crowd-sourced labels to more expert classification systems like the Hubble tuning fork is essential. Multiple ResNet50 models are trained on the crowd-sourced Galaxy Zoo 1 and 2 datasets as well as the expertly labelled EFIGI catalogue to classify galaxies according to their Hubble types. To study the generalisation of the models trained on crowd-sourced data against the models trained on expert data, the expert Revised Shapley-Ames catalogue is used an unseen test set. Deep Metric Learning techniques are used to fine-tune classification models to improve on the current state-of-the-art results for classifying galaxies. The results show that Transfer Learning coupled with the ResNet50 outperforms self-defined rules for galaxy classification, indicating the effectiveness of machine learning for galaxy classification. The results further demonstrate an improvement on the current state-of-the-art accuracy for both the Galaxy Zoo 2 and EFIGI data, using Transfer Learning with the ResNet-50 model. The mean average precision values for both the crowd-sourced and expert models indicated that the models are comparable. However, confusion matrices reveal that the models trained on the expert dataset outperformed the models trained on the crowd-sourced data in terms of actual vs. predicted labels. The result highlights the need for caution when utilising crowd-sourced labels. The results further show that a model that has been pre-trained on crowd-sourced data using Label Smoothing Cross-Entropy can be fine-tuned using Deep Metric Learning to achieve the state-of-the-art performance in galaxy morphology classification. Finally, Transfer Learning from crowd-sourced labelled data to expert-labelled data leads to significant improvement in classification accuracy.

Acknowledgements

The continued time, guidance and support from my supervisors, Prof. Terence Van Zyl and Dr. Matthew Woolway, helped me not only achieve a deep understanding of the research topic and the machine learning techniques utilised in this research, but also served to develop strong, independent research skills. Under the guidance of two experts, I feel confident that I have gained invaluable knowledge and skills which I hope to continue developing in the future.

I would be remiss to not mention the emotional support provided by my parents whose unwavering support and belief in me served as strong motivation for the duration of this research.

—

Contents

List of Figures

List of Abbreviations

CNN	**C**onvolutional **N**eural **N**etwork
RSA	**R**evised **S**hapley-**A**mes
GZ1	**G**alaxy **Z**oo **1**
GZ2	**G**alaxy **Z**oo **2**
GZ2-H	**G**alaxy **Z**oo **2** **H**ubble types
MSE	**M**ean **S**quared **E**rror
RMSE	**R**oot **M**ean **S**quared **E**rror
LS-CE	**L**abel **S**moothing **C**ross-**E**ntropy
P-Anchor	**P**roxy-**A**nchor
P-NCA	**P**roxy **N**eighbourhood **C**omponents **A**nalysis
Norm-S′max	**N**ormalised **S**oftmax
MAP	**M**ean **A**verage **P**recision
MAP@R	**M**ean **A**verage **P**recision **A**t **R**
ROC	**R**eceiving **O**perating **C**haracteristic Curve
DML	**D**eep **M**etric **L**earning
Class	**C**lassification

1 Introduction

1.1 Background

A galaxy is defined as a huge system, bound by gravity, that consists of stars, stellar remnants, an interstellar medium of gas and dust, and dark matter. A method that is used to separate galaxies into classes based on their physical structures is called galaxy morphological classification. Galaxies have three main classifications, namely elliptical (which is described by its ellipse-shaped light profile), spiral (which is disk-shaped and has multiple curved arms that originate from it) and irregular (which is any galaxy that does not fit into the first two categories). In studying the origin and evolution of the universe, one area of focus is galaxy classification, as the type and formation of galaxies often offers clues and insights into the origin and evolution of the universe (Khalifa et al., 2017). An academically accepted system (Khalifa et al., 2017; Barchi et al., 2020) for classifying galaxies is the Hubble tuning fork.

The majority of galaxy classifications available in catalogues come from human experts inspecting images of galaxies and manually labelling them. With an astronomical number of images of galaxies available, in the order of 1×10^6, it becomes increasingly laborious for scientists to classify these galaxies manually (Cai et al., 2020). As a result of the increased availability of images of galaxies, citizen science and machine learning are popular methods for classifying galaxies (Cai et al., 2020; Lintott et al., 2011). If citizen science and machine learning are to be coupled to automate galaxy classification, it is essential to understand how well the crowd-sourced labels generalises to an expertly labelled catalogue like the Revised Shapely-Ames (RSA) catalogue (Sandage and Tammann, 1981).

1.1.1 Image Classification

The main area of focus in this research is to use machine learning for classifying images. Image classification is referred to as a supervised learning problem. In image classification, a set of target classes (objects to identify in images) is defined, and a model is trained to recognize the target classes using example photos for which we know the labels. Convolutional Neural Network (CNN) architectures are highly effective for image classification (Yamashita et al., 2018) and are discussed in-depth in Section 2.2.1.

1.1.2 Research Problem

Due to the increased availability of images of galaxies, in the order of 1×10^6 (Cai et al., 2020), citizen science and machine learning are popular methods for classifying galaxies (Cai et al., 2020; Lintott et al., 2011). This dissertation reports on the results and findings from the experiments in which various machine learning techniques were used for galaxy classification. First, a comparison is drawn between a set of self-defined rules and machine learning techniques for galaxy classification to test the effectiveness of machine learning techniques for galaxy classification. Second, a comparison is drawn between the generalisation of the crowd-sourced GZ2 data (Hart et al., 2016) vs. the expertly labelled EFIGI catalogue (Vaucouleurs et al., 2013) to the Hubble tuning fork. Finally, Deep Metric Learning techniques are utilised to fine-tune classification models to improve on the current state-of-the-art in galaxy classification.

1.1.3 Literature Review

Previous work in this field has shown promising results across various image processing and machine learning techniques. (Khalifa et al., 2017) describe a CNN used for galaxy classification. The EFIGI catalogue was used to train the CNN. The first layer of the CNN is the input layer into which the images are input. The second layer of the CNN is the convolutional layer (for feature extraction). Rectified Linear Unit (ReLu) serves as the non-linear activation function of the third layer. In the fourth layer, pooling is performed. The fifth layer is a fully-connected layer with 24 neurons, and also uses a ReLu activation function. The output layer consists of three neurons and makes use of a soft-max activation function to determine class membership. Classifying three classes of galaxies (elliptical, spiral and irregular), they achieved an accuracy of 97.27%. In Section 3.3 we show that we were able to marginally outperform their 3-class EFIGI classification.

(Calleja and Fuentes, 2003) describe an approach in which they propose a method of using image analysis, feature extraction using principal components analysis (PCA) and machine learning for galaxy classification. The data used consists of 310 images of galaxies that have different sizes, colour formats and it was found that the galaxy in each image is not always found in the centre of the image. The methodology followed can be summarised into three steps: image analysis module (IAM), data compression module (DCM) and machine learning module (MLM). The IAM step binarises each image (i.e makes it black and white) and ensures that the main axis of each image is horizontal by rotating each image so, in other words, this process standardises each image. In the DCM step, PCA is used to extract the most important features of each image. In the MLM step, they trained an artificial neural network (ANN) and a locally weighted regression (LWR) method on their data. They tested the models on varying classes of galaxies $(2, 3, 5, 7)$ and it was found that the LWR method classifying two classes of galaxies (elliptical and spiral) yielded the

best accuracy at 95.11%. This accuracy dropped to 91% when classifying galaxies into 3 classes (elliptical, spiral and irregular), 52.29% when classifying galaxies into 5 classes and 44.63% when classifying galaxies into 7 classes.

Another approach by (Kasivajhula et al., 2003) outlines a process of image processing, feature extraction and, finally, training of 3 different machine learning models on the extracted features. Their dataset consisted of 119 images obtained from Zsolt Frei's galaxy catalogue (Frei et al., 1995). To process the images they binarised them using a pre-determined threshold to remove faint, extraneous noise. The two principal components of each image were then calculated and used to rotate the main axis of the galaxy. Each image was then scaled to a uniform size of 128×128 pixels. 6 morphic features (which are based on the perceived visual characteristics of the galaxy) were then calculated for each image: elongation, form-factor, convexity, bounding-rectangle-to-fill-factor, bounding-rectangle-to-perimeter and asymmetry index. For classification, the authors trained a support vector machine (SVM), random forest (RF) and a naive Bayes classifier (NB) and tested their models on 3 and 7 classes of galaxies. They tested the algorithms using the morphic features only, the principal components only and a combination of the morphic features and principal components and found that using the RF with only morphic features yielded the best results, classifying galaxies into 3 classes with an accuracy of 85.72%. This accuracy dropped to 30.03% when classifying galaxies into 7 classes, also using a comnination of morphic features and principal components.

Galaxy Zoo became the focus of a Kaggle competition in which pariticipants were supplied with 61578 images from the Galaxy Zoo 1 dataset to train machine learning models to predict the 37-class response vectors. (Dieleman et al., 2015) describe the winning solution using a CNN. In the pre-processing step, they cropped the images from the original 424×424 size to 207×207 and then downscaled them 3 times to 69×69. To avoid loss of information, they used an object detection algorithm to estimate both the Petrosian radii and positions of the objects, which they use to centre and rescale all images to standardise the sizes of the objects. To synthetically increase the number of training samples, each training image was randomly perturbed in five ways: rotation, translation, scaling, flipping and brightness adjustment. Then, they extracted multiple 'viewpoints' from each image by rotating one image multiple times. The best performing model was obtained by averaging over 60 transformations of the input data and averaging 17 networks and reported the root mean squared error (RMSE) value to be 0.0749.

The work by (Gauthier et al., 2016) describes a method of galaxy classification on the Galaxy Zoo 1 dataset. They start with a few image processing techniques to clean the data. As target variables, they converted the Galaxy Zoo 1 response vectors to five classes (namely elliptical, spiral, round, disk and other) using a self-defined rule based on the decision tree in Figure 2.6. For the image processing, first, they crop the images, to reduce the file size and remove many background secondary objects.

They then rotate each image so that the principal axis of each galaxy is vertical. Finally, to remove background noise they carve out isophates at 10% of the maximum intensity threshold of each image and set the pixel values of all non-central isophotes to zero. They took all the processed images and applied 125-component PCA on them (they chose 125 components because it retained over 99% of the variance). They trained several machine learning models on the principal components (an RF, AdaBoost, K-Nearest Neighbours (KNN), Decision Tree and SVM) and found that the RF produced the best accuracy of 67%, using 5 classes of galaxies (spiral, elliptical, round, disk or other).

(Lukic and Bruggen, 2016) replicated the Galaxy Zoo Kaggle competition's winning solution by training a CNN architecture which had varying convolutional and pooling layers. The ReLU and Parametric Rectified Linear Unit (PReLU) activation functions were both used in their experiments. The stochastic gradient descent (SGD) optimisation algorithm and a minibatch size of 16 were used to train the networks. The models were trained over 200 epochs. The initial learning rate was 4×10^{-2}, which at epoch 150 decreased to 4×10^{-4}. Their results showed that by using three convolutional layers with the PReLU activation function produced lower training and validation errors compared to the ReLU function using the mean squared error (MSE) loss for training. Using the PReLu activation function, the RMSE on the training set was reported to be 0.4111, and the RMSE on the validation set was reported to be 0.4062. The difference in the RMSE value reported by (Lukic and Bruggen, 2016) and (Dieleman et al., 2015) can be attributed to the difference in CNN architectures used as well as the advanced image pre-processing techniques used by (Dieleman et al., 2015).

(Walmsley et al., 2019) discuss that the shortcoming of using CNNs in approximating human classifications is that uncertainty is not accounted for in training labels, which limits the ability of the networks to effectively learn from all available data. To solve this, they propose a method of using Bayesian CNNs and a novel generative model to infer posterior probability distribution functions for images of galaxies in the GZ1 dataset. The Bayesian CNN can be trained using galaxy images with uncertain labels and then predict the probability of each possible label for unlabelled images of galaxies. They then use an Active Learning approach to show how these posterior probability functions can be used to choose galaxies with the most information for labelling. They show that through the use of Active Learning, selecting the galaxies with the most information, combined with the use of a Bayesian CNN, they would require somewhere between 35% and 60% fewer galaxies with labels for training, depending on the morphological feature being used.

(Dai and Tong, 2018) propose a method of combining a variant of residual networks with other popular CNNs to classify a sample of the GZ1 dataset into five classes: completely round smooth, in-between smooth, cigar-shaped smooth, edge-on and spiral. Their results showed that across various metrics such as accuracy, precision,

recall, F_1 value and Area Under Curve (AUC) their model was in line with the state-of-the-art classification performance of other approaches. The overall accuracy of their model on the testing set was reported to be 95.21%.

A Capsule Network is a neural network architecture which was introduced after the conclusion of the Galaxy Zoo 1 project (Katebi et al., 2018). Capsule Networks are rotationally invariant and spatially aware (Katebi et al., 2018). (Katebi et al., 2018) trained a Capsule Network using the Galaxy Zoo 1 response vectors. On the unseen test set, the Capsule Network produced an RMSE value of 0.1030. They then tried an approach in which they used the answer with the highest number of responses as the label of the sample and trained a Capsule Network classifier. The classifier only used responses to the first question in the Galaxy Zoo decision tree 2.6 for training, “Is the galaxy simply smooth and rounded with no sign of a disk?". The question has 3 responses, “round and smooth" (i.e. elliptical galaxies), “objects with disks" (spiral galaxies) and “artifact or star". These 3 responses were used as the target classes for training. The Capsule Network achieved an accuracy of 98.77% classifying galaxies into these 3 classes, as opposed to the accuracy of 96.96% achieved by the baseline model.

(Barchi et al., 2020) provide a comparative study of machine and deep learning applied to galaxy morphology classification. They use the Galaxy Zoo 2 Hubble types (GZ2-H) dataset in their study. Classifying galaxies into 2 classes, the machine learning algorithms produced an accuracy of 98% vs the 94% obtained when employing a CNN. This accuracy dropped to 65.20% when classifying galaxies into 11 Hubble types. In Section 3.3 we show that we were able to significantly outperform the accuracy they obtained classifying 11 Hubble types.

(Boudiaf et al., 2020), described a set of experiments in which they showed that combining Label Smoothing with Cross-Entropy for classification is able to outperform state-of-the-art Deep Metric Learning (DML) methods. In their experiments, they used various benchmark classification datasets to test their models, such as the Stanford Cars Dataset (Krause et al., 2004), the Caltech-UCSD Birds-200-2011 (CUB) Dataset (Wah et al., 2011) and others. In Section 3.4 we show that models trained using Label Smoothing Cross-Entropy can be fine-tuned using DML techniques to outperform the state-of-the-art in galaxy classification.

(Cai et al., 2020) 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, and a variational autoencoder (VAE) which compresses the activation maps into a 1-D vector, and a classifier that generates labels from the activation maps. The dataset was made up of 35784 images generated from 36 N-body simulations and is available

online. In Section 3.4 we show that we trained a ResNet50 model on their simulated dataset, but that the model could not transfer to our problem of predicting the Hubble types of galaxies.

1.2 Significance and Motivation

With the increase in the number of images of galaxies being made available, in the order of 1×10^6 (Cai et al., 2020), as well as the increase in popularity of machine learning techniques in various industries, researchers have turned to citizen science and machine learning to automate the process of galaxy classification (Lintott et al., 2011; Hart et al., 2016).

Although some work has already been performed in the field of galaxy classification, we have shown in Section 1.1.3 that there has been no attempt to investigate how well crowd-sourced labels generalise to a more expert system like the Hubble tuning fork.

1.3 Research Aims and Objectives

This research explored the effectiveness of various machine learning techniques for galaxy classification. It is also investigated how well crowd-sourced labels of galaxies could generalise to the expert Hubble tuning fork classification system. Finally, the use of DML techniques to fine-tune classification models to outperform the state-of-the-art in galaxy classification was investigated.

1.4 Research Questions

How well does a model trained using crowd-sourced labels generalise to the expert Hubble tuning fork classification system? Furthermore, can a classification model be fine-tuned using Deep Metric Learning techniques to outperform state-of-the-art in galaxy classification?

1.5 Delineations, Limitations and Assumptions

It is assumed that there is only one galaxy in each image, leaving room for further work to include an object detection algorithm to detect and classify multiple galaxies in one image. If the galaxy isn't in the center of the image, or if part of it is cut off, it is assumed that the model will still try to identify patterns in the image and if it can't then it should classify the galaxy as the Irr class.

1.6 Contributions of the dissertation

[21] Mohamed Z. Variawa, Terence L. van Zyl, and Matthew Woolway. "A rules-based and Transfer Learning approach for deriving the Hubble type of a galaxy from the Galaxy Zoo data". In: *23rd International Conference on Information Fusion (FUSION 2020)* (2020). DOI: IEEEcatalognumber : CFP18FUS - ART , ISBN : 978-0-964527-6-2.

[22] Mohamed Z. Variawa, Terence L. van Zyl, and Matthew Woolway. "Comparing Generalisation Using Crowd-sourced vs Expert Labels for Galaxies Classification". In: *7th Intl. Conference on Soft Computing & Machine Intelligence (ISCMI 2020)* (2020).

[23] Mohamed Z. Variawa, Terence L. van Zyl, and Matthew Woolway. "Transfer Learning and Deep Metric Learning for Automated Galaxy Morphology Classification". In: *Neural Computing & Applications* (Submitted 2021).

1.7 Outline

This Section provides a brief outline of the structure of this dissertation.

Chapter 2 presents a detailed description of the methodology used to answer the research question. Section 2.3 explores the deep learning architecture, CNN, and techniques, Transfer Learning and Deep Metric Learning, used in this research. An analysis of the data used in this research is provided in Section 2.3.2. Finally, in Sections 2.4, 2.5 and 2.6 the experimental setup is discussed in depth.

Chapter 3 discusses the results from the experiments. Sections 3.2, 3.3 and 3.4 presents the results of the experiments in the form of tables, confusion matrices and Receiver Operating Characteristic curves as well as provides a comparison with the current state-of-the-art in galaxy classification. A detailed discussion around the findings is provided in Section 3.5.

Chapter 4 summarises the main points from this dissertation, provides a summary of the findings and provides suggestions on future work which may improve on the work in this dissertation.

2 Research Methodology

2.1 Introduction

In this Chapter the experimental setup used to answer the research question is discussed. Section 2.2.1 provides an introduction to some of the necessary theory relevant to this dissertation such as the deep learning architecture, specifically Convolutional Neural Networks, used in this research. The deep learning techniques Transfer Learning and Deep Metric Learning which are discussed in Sections 2.2.2 and 2.2.3, respectively. A description of the software and hardware used is provided in Section 2.3.1. In Section 2.3.2 a description of each of the datasets used in this research is provided and in Section 2.3.3 an analysis of the datasets is provided. In Sections 2.4, 2.5 and 2.6, the experimental setup is described in depth.

2.2 Theory

2.2.1 Neural Networks and Convolutional Neural Networks

An artificial neural network (ANN) is a computational system that is inspired by the human brain. These systems learn to carry out tasks by considering training samples, without being pre-programmed with any specific rules (Wu, 1992). The general structure of an ANN is illustrated in Figure 2.1, and can be described as:

- Artificial Neuron: Each node in each layer of the ANN is known as an artificial neuron, which is just a placeholder for certain values. For example, the input layer has neurons which store the input data.
- Input layer: The layer into which we read the input data to be transformed by the subsequent layers.
- Hidden layer(s): The layer(s) between the input layer and the output layer (if there are multiple hidden layers, we call it a deep neural network). The artificial neurons in this layer receive as input a set of weighted values; the output of a function (called the activation function) which takes as input the dot product between the values in the neurons of the previous layer with their weights.
- Output layer: The final layer of artificial neurons from which we get an output.

- Weights: The connections between the neurons in each layer are known as weights.

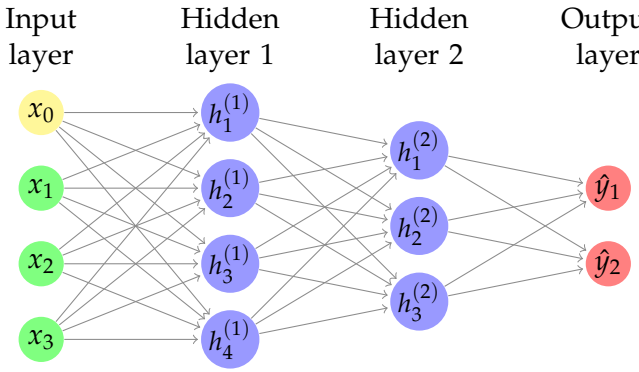

FIGURE 2.1: Structure Of An ANN With 4 Input Nodes, 2 Hidden Layers and 2 Output Nodes

An activation function uses an input (or set of inputs) to produce an output, for a particular node. The Sigmoid function is an example of an activation function widely used in classification problems:

$$Sigmoid(x) = \frac{e^x}{1 + e^x}$$

The process of passing a set of inputs, into the input layer, and calculating their dot product with their weights and then passing that dot product into an activation function to get the values of the hidden layer is known as forward propagation (Wu, 1992). The same process is then applied to the values in the hidden layer to get the output and comparing that output to the actual output, as illustrated in Figure 2.2:

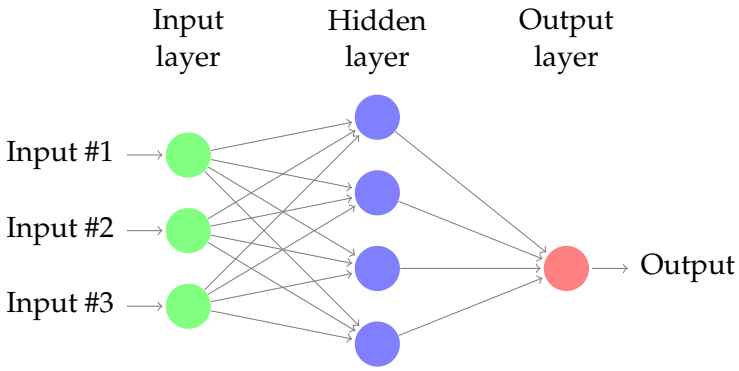

FIGURE 2.2: Forward Propagation

For the network shown in Figure 2.2, if the initial weight matrix shared between the input and hidden layers is given by Table 2.1 and the initial weight vector shared between the hidden and output layers is given by $[0.2, 0.3, 0.6, 0.1]$, then for the input $X = [0.1, 0.4, 0.5]$, the output of the network in Figure 2.2 would be 1.

$$\begin{bmatrix} 0.1 & 0.2 & 0.4 & 0.3 \\ 0.5 & 0.4 & 0.7 & 0.9 \\ 0.2 & 0.6 & 0.3 & 0.8 \end{bmatrix}$$

TABLE 2.1: Example Initial Weights Between Input And Hidden Layers

In classification problems, the Cross-Entropy loss function can be used to compare the predicted output of the model to the true output value. As the predicted probability diverges further from the actual output, the Cross-Entropy loss increases. For example, if the predicted probability is 0.1 and the actual output is 1, the loss value would be high. A model that has been fit perfectly would have a loss of 0. The Cross-Entropy loss function is given by:

$$L = -\sum_i \left(y_i' log(y_i) + (1 - y_i') log(1 - y_i) \right),$$

where y_i' represents the true value and y_i represents the predicted value.

Backpropagation is a method used to train the ANN. In the backpropagation method, the predicted output and the true output are used to calculate the error (loss) of the output layer. The error of the output layer is sent through the network backwards (output layer $\rightarrow$ hidden layer(s) $\rightarrow$ input layer) and is used to update the weights in each layer. The weight update rule is given by:

$$W_i^* = W_i - a \times \left(\frac{\partial Error}{\partial W_i} \right),$$

where W_i is the old weight, W_i^* is the updated weight, a is the learning rate and $\frac{\partial Error}{\partial W_i}$ is the partial derivative of the error (loss) function with respect to weight W_i.

An ANN architecture that is very effective in image classification is the CNN (Yamashita et al., 2018). The structure of a CNN, specifically the ResNet50 architecture used in this research, is shown in Figure 2.3:

Features are extracted from an image in the convolutional layer. By using small grids of the input data to learn the features of images, the spatial relationship between neighbouring pixels is maintained in the convolutional step. In the convolution step, we use a matrix which we call the filter to extract features from the input image (Yamashita et al., 2018). We then get the dot product of this filter with each $n \times n$ (where n is the size of the filter) grid within the original input image and produce a matrix which we call the feature map (Yamashita et al., 2018). There are different types of filters for different operations such as edge detection. In practice, a CNN should learn the values of the filter matrix by itself during training (however we still need to specify some information such as how many filters we want to use, the size of the filters, the network architecture etc. before training). We can improve the networks ability to recognize patterns in unseen images by using more filters

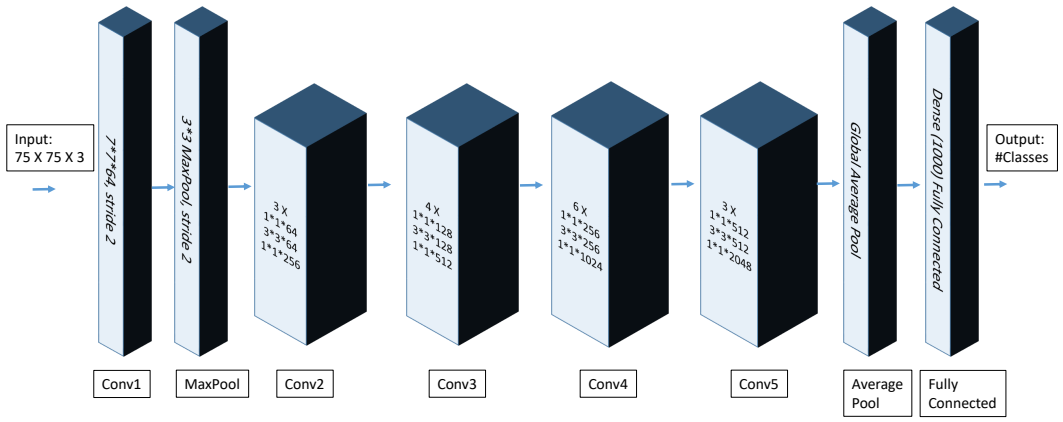

FIGURE 2.3: ResNet50 Architecture

which in turn leads to an increase in the amount of features the network can extract (Yamashita et al., 2018). The three parameters that are used before training which determine the size of the feature map:

- Depth: This parameter describes how many filters the convolution operation will use.
- Stride: This parameter describes how many pixels in the input matrix the filter matrix will slide over in one go. For example, with a stride of 1 the filters will move over one pixel in input matrix at a time.
- Zero-padding: Filling (padding) the border of the input matrix with zeros, this ensures the filter can be applied to the outer (border) elements of the image matrix. Zero-padding enables us to determine the size of the feature maps.

The convolution process is linear where most real-world data is non-linear, we therefore may want to introduce some non-linearity to our network after the convolution process. This can be achieved through the use of non-linear activation functions such as the ReLU function, which is given as

$$Output = Max(0, Input)$$

ReLU is applied per value in the matrix (i.e per pixel in the image), where every pixel in the feature map with a negative value is changed to zero.

To retain the most important information of the feature map while its dimensionality (size) is reduced, spatial pooling is applied (Yamashita et al., 2018). Spatial pooling can vary between Max Pooling, Average Pooling, Sum Pooling etc. Given a spatial

neighbourhood (e.g. a 2×2 window), Max Pooling would output the largest number from the feature map within the 2×2 window. Average Pooling would output the average of all the numbers within the 2×2 window. After performing pooling, a method known as flattening is applied to the pooled matrices (Yamashita et al., 2018). Flattening transforms a 2-D matrix into a 1-D vector.

The fully-connected layer is a 'traditional' ANN, as described above. The term 'fully-connected' implies that there is a connection between each neuron in adjacent layers. The convolutional and pooling layers produce output that represents high-level features of an image which are then used by the fully connected layer to classify the image into various classes based on patterns learned from the training data. The CNN is then trained using the backpropagation process described above.

The CNN architecture used in this research is the ResNet50 (Hira et al., 2020). In the ResNet50 architecture, residual functions in respect of layer inputs are used to reformulate network layers. The ResNet (residual network) solves the vanishing gradient problem by introducing skip connections (Hira et al., 2020). The ResNet50 architecture allowed for the development of deeper CNNs without compromising on accuracy (Hira et al., 2020). ResNet50 uses batch normalization (Hira et al., 2020). The architecture is shown in Figure 2.3.

2.2.2 Transfer Learning

Although deep learning is effective in making predictions (especially in image classification) (Weiss et al., 2016), training deep learning models requires a significant amount of time, computational resources and data (Weiss et al., 2016). Some of these limitations can be circumvented through the use of Transfer Learning. In Transfer Learning, a deep learning model that has been trained on one dataset serves as the initialisation for training the model on another dataset (Pan and Yang, 2010). The two models may be from the same or similar domains (Pan and Yang, 2010). It has been shown that Transfer Learning is highly effective in image classification problems (Weiss et al., 2016) and has been used extensively in this research.

2.2.3 Deep Metric Learning

Machine learning classification problems involve passing training data through a neural network and having the final layer output the class probabilities. This leads to the requirement of having a large training dataset in which the training samples have similar characteristics to the samples we expect when testing the model.

DML involves training a neural network to learn a complex, non-linear mapping of high-dimensional input data to a lower-dimension manifold, known as an embedding space, using pairwise distances (e.g. the Euclidean norm or cosine similarity) as a metric (Boudiaf et al., 2020). A deep encoder, trained on higher-dimension input data (e.g. images), maps similar samples (i.e. embeddings with a low Euclidean

norm) together within the embedded space. The loss functions used to train the deep encoder are called pairwise losses. During training, the pairwise losses penalise the model so that it encourages a smaller distance for training samples sharing the same class and a larger distance for training samples with different classes (Boudiaf et al., 2020).

A Siamese Network is an architecture consisting of two or more parallel neural networks that share the same weights, where each network takes as input a different sample and the outputs are compared to provide a measure of similarity between the inputs (Hadsell et al., 2006). Siamese Networks allow for the implementation of DML.

In image classification problems, the identical neural networks may be CNN's. Using the training images, the model learns to effectively extract latent features from the input images, in the form of 1-D feature vectors (i.e. embeddings). In order to classify a new image, the trained network can be used to produce the embeddings for the images in the training set as well as the embedding for the new image. A distance function is used to compute the distances between the embedding for the new image and the embeddings for the training images. The new image will then be classified according to the class of the training image it is closest to (i.e. its nearest neighbour). To test the information retrieval accuracy of the model, the Mean Average Precision Value (MAP) can be calculated. The MAP@R value is a measure of how many of the top R nearest neighbours share the same class as the new image to be classified. The structure of the Siamese Network used in this research is shown in Figure 2.4 where the ResNet50 model architecture is shown in Figure 2.3.

The sampling methods for training DML models are the triplet and pair mining functions (Boudiaf et al., 2020):

- In triplet mining, each training sample, also known as the anchor (A), is fed into the network along with two other samples: A positive sample (P), whose class/label will be the same as the anchors', and a negative sample (N), whose class/label will be different to the anchors'.
- In pair mining, samples are fed into the network 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 be different from the anchors'.
- During training, the pairwise loss function penalises the network so that the distance (e.g. Euclidean norm) between the anchor and positive embedding spaces will become smaller 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 become larger at

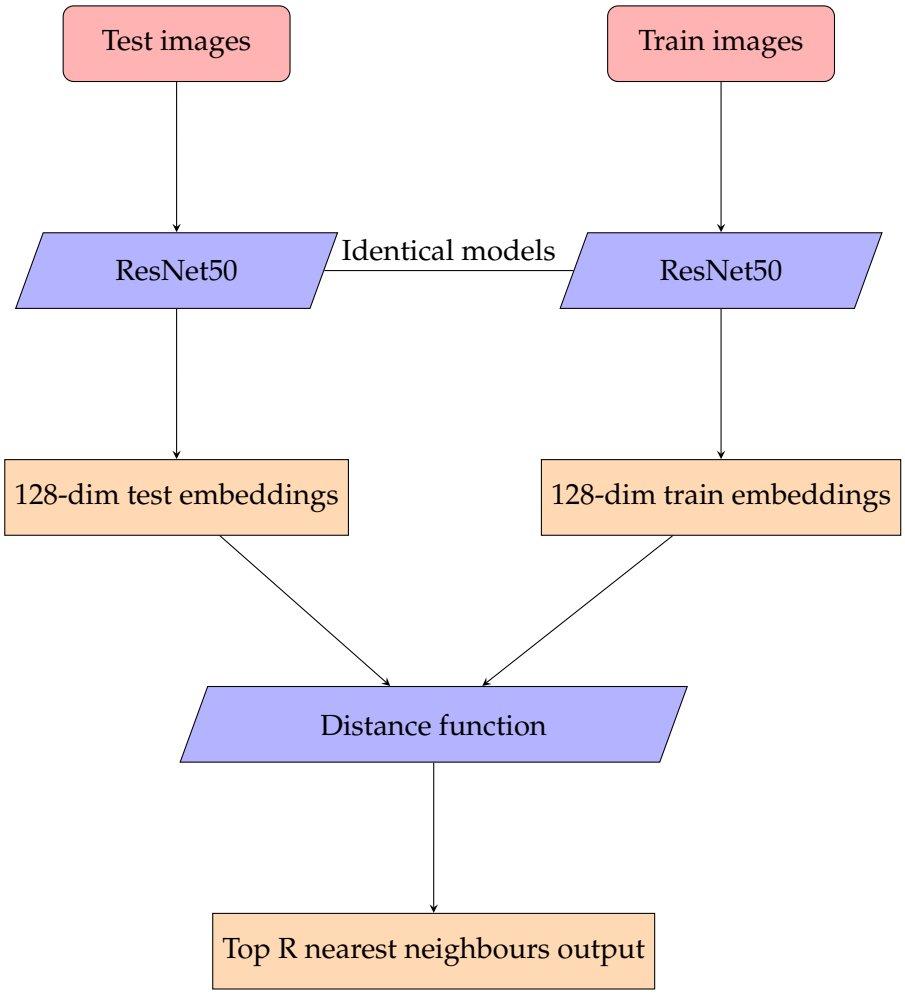

FIGURE 2.4: Siamese Network Structure

each iteration (i.e. the embeddings for samples in different classes will move further away from one another).

2.3 Research Design

2.3.1 Research Instruments

The programming language was Python 3.7 and the models were trained using PyTorch version 1.6.0. The CPU used was an Intel® Core™ i7-7700 CPU @ 3.60 GHz. The computer has 16GB DDR4 2400 MHz RAM and an Nvidia® GeForce® GTX 1060 6GB GDDR5 GPU. The code for all the experiments can be found on GitHub (Variawa et al., 2020c).

2.3.2 Data

This Section is opened by describing an academically accepted (Khalifa et al., 2017; Barchi et al., 2020) system for labelling galaxies called the Hubble tuning fork. The citizen-science Galaxy Zoo project and the crowd-sourced Hubble labels derived

therein are then described. The expertly-labelled Revised Shapley-Ames and EFIGI catalogues are also described in this Section.

The Hubble tuning fork classification scheme

An academically accepted (Khalifa et al., 2017; Barchi et al., 2020) system for classifying galaxies based on their physical structures, the Hubble tuning fork classification scheme, is shown in Figure 2.5. The tuning fork is considered robust as it accounts for elliptical structures and both barred, and non-barred spiral structures (Khalifa et al., 2017; Barchi et al., 2020).

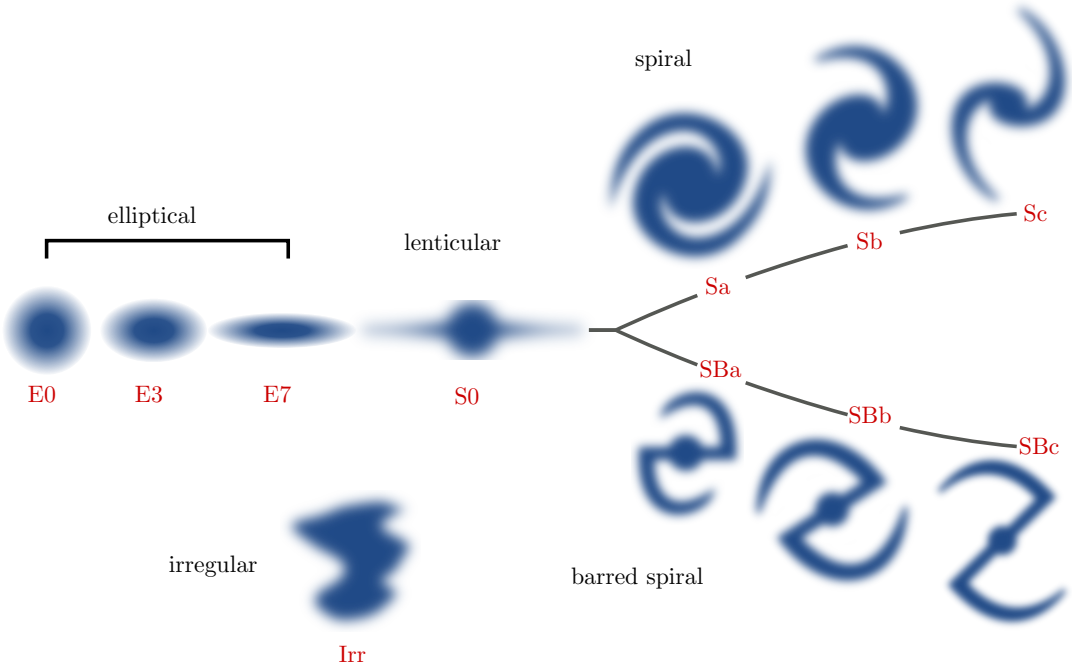

FIGURE 2.5: Hubble Tuning Fork Used In This Research For The 9 And 11 Hubble Classes

The Galaxy Zoo project

GZ1 is a citizen-science galaxy classification project described by (Lintott et al., 2011). The project aimed to solve the problem of astronomers spending large amounts of time manually classifying images of galaxies (Lintott et al., 2011). A set of questions (11 in total) about the structure of a galaxy in an image was presented to participants (members of the public). There were 37 possible responses as every question had a fixed number of responses. Question 1 "Is the object a smooth galaxy, a galaxy with features/disk or a star?" had the following three options: smooth, features or disk, star/artefact. Participant responses for each question were counted and assigned a probability of being chosen based on the collated responses. Table 2.2 shows example responses for the first question.

TABLE 2.2: GZ1 Example Probabilities

Class1.1	**Class1.2**	**Class1.3**
0.3831	0.6169	0

Table 2.2 shows that (for question 1) that 61.68% of people chose the second response (a galaxy with features/disk) which indicates that Class1.2 had a probability of 0.6169.

The questions posed by (Lintott et al., 2011), with their number of possible responses, are provided below:

- Q1. Is the object a smooth galaxy, a galaxy with features/disk or a star? *3 responses*
- Q2. Is it edge-on? *2 responses*
- Q3. Is there a bar? *2 responses*
- Q4. Is there a spiral pattern? *2 responses*
- Q5. How prominent is the central bulge? *4 responses*
- Q6. Is there anything "odd" about the galaxy? *2 responses*
- Q7. How round is the smooth galaxy? *3 responses*
- Q8. What is the odd feature? *7 responses*
- Q9. What shape is the bulge in the edge-on galaxy? *3 responses*
- Q10. How tightly wound are the spiral arms? *3 responses*
- Q11. How many spiral arms are there? *6 responses*

Because a question had multiple possible responses, particular responses to questions might lead to different subsequent questions. A decision tree, presented in the Galaxy Zoo publication (Lintott et al., 2011), has been compiled to show the all different paths between questions, as shown in Figure 2.6.

Galaxy Zoo 2 (GZ2), introduced by (Hart et al., 2016), is the second phase of GZ1. GZ2 extended the original GZ1 classification to include more information on each galaxies' structure, such as bars, spirals, bulges, and others (Barchi et al., 2020). The GZ2 dataset utilised the Sloan Digital Sky Survey website (SDSS) to obtain 239573 images of galaxies. The 37-class responses in the GZ2 were accompanied by the Hubble type (GZ2-H) for each galaxy (derived from participants' responses) (Barchi et al., 2020; Hart et al., 2016).

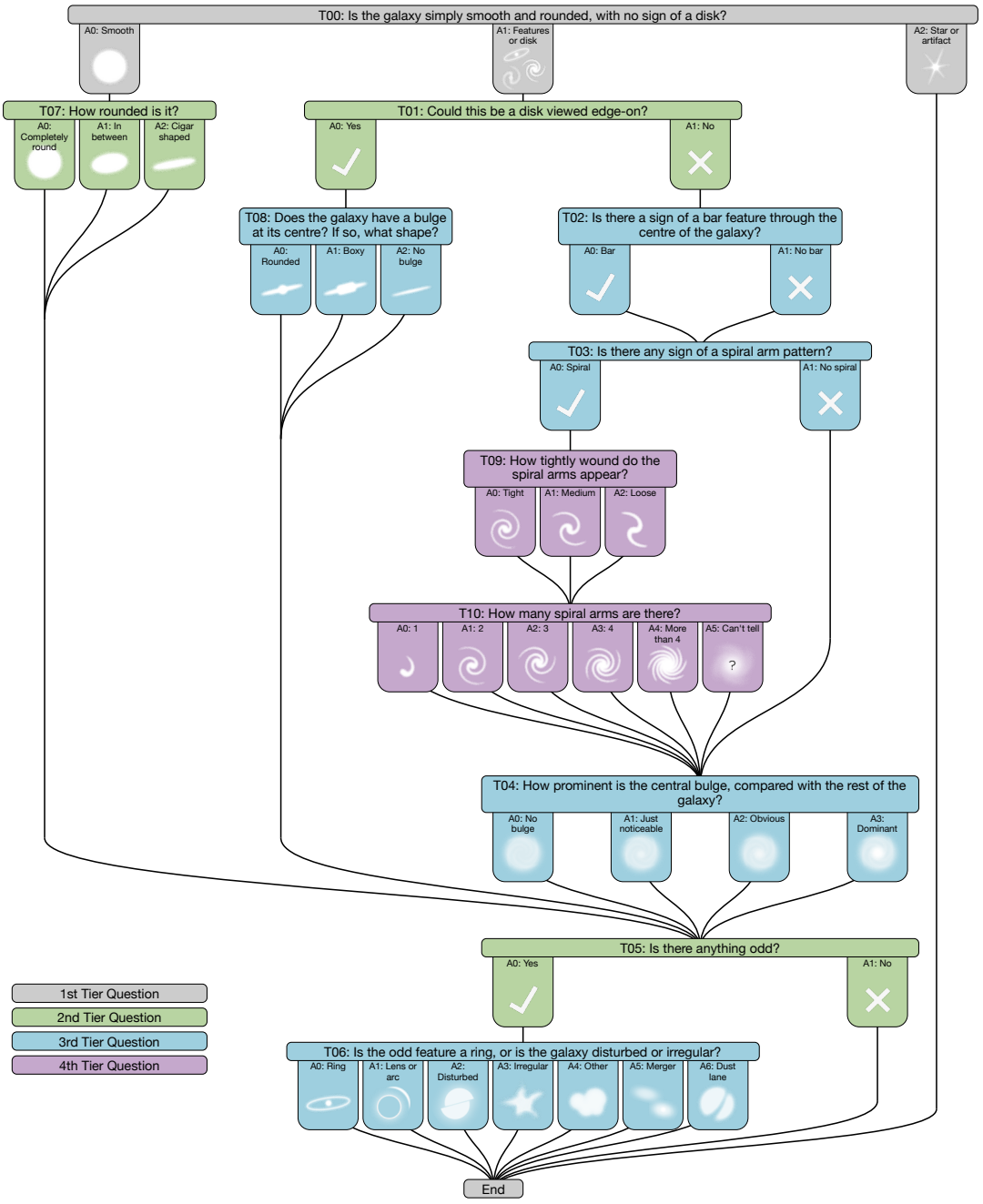

FIGURE 2.6: The Galaxy Zoo Decision Tree (Lintott et al., 2011)

The Revised Shapley-Ames catalogue of bright galaxies and the EFIGI catalogue

The Revised Shapley-Ames catalogue of bright galaxies contains 1249 images of galaxies with their corresponding expert Hubble types (Sandage and Tammann, 1981). The images in the catalogue were manually sourced from the SDSS website, as the set of images is not available.

A subset of 4488 images of galaxies in the *Third Reference Catalogue of Bright Galaxies* (Vaucouleurs et al., 2013) is available in the EFIGI catalogue. The galaxies in the EFIGI catalgoue are classified according to the de Vaucouleurs system which is an extended alternative to the Hubble tuning fork (Vaucouleurs et al., 2013). The de

Vaucouleurs system can be mapped back to the original tuning fork (Vaucouleurs et al., 2013). The catalogue does not specify specific elliptical types (Vaucouleurs et al., 2013); therefore, all ellipticals are grouped into type E, as shown in Table 2.4.

The image pre-processing techniques applied to the Galaxy Zoo 2, EFIGI and RSA data were: First, cropping the image to a size of 224×224 followed by downscaling the image to a size of 75×75. Random horizontal and vertical flips (i.e. augmentation techniques) are then applied to the images. Finally, the data is normalised.

2.3.3 Analysis

Table 2.3 summarises the information on the data used to train and test our models.

TABLE 2.3: Datasets Metadata

Paper	Dataset	# Samples	# Hubble Classes	Train/Test/Validation Split (%)
(Lintott et al., 2011)	GZ1	61578	-	70 / 15 / 15
(Hart et al., 2016)	GZ2-H	239573	9, 11	70 / 15 / 15
(Vaucouleurs et al., 2013)	EFIGI cat.	4488	9	70 / 15 / 15
(Sandage and Tammann, 1981)	RSA cat.	1249	9, 11	Used as an unseen test set

Table 2.4 provides a consolidated view of the percentages of each of the Hubble classes shown in Figure 2.5 for each dataset. Table 2.4 shows the classes Irr, SBa and Sa are severely under-represented in the crowd-sourced GZ2-H data, a problem which is not present in the expert EFIGI and RSA catalogues.

TABLE 2.4: 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

Cross-referencing the RSA galaxies with the EFIGI and GZ2-H sets showed that 420 galaxies were in both the RSA and EFIGI sets. To ensure that no galaxies from the RSA catalogue would be used during training, the overlapping galaxies were removed from the EFIGI set before training, with 4068 images remaining in the EFIGI catalogue.

The models trained on the 37-class vectors in the GZ1 and GZ2 datasets were regression problems, for which the metric used is the RMSE. The classification models are assessed using accuracy, which is common in the literature. The metrics used to evaluate the DML models are accuracy, and MAP and MAP@R values.

Confusion matrices and Receiver Operating Characteristic (ROC) curves are used to further investigate the predictions as well as to compare the performance between the various models.

2.4 Effectiveness of machine learning techniques for galaxy classification methodology

The work described in this Section has been published in the proceedings of the *23rd International Conference on Information Fusion* by (Variawa et al., 2020a).

The initial set of experiments focused on comparing the effectiveness of machine learning techniques against a set of self-defined rules for galaxy classification. The dataset used for training was taken from the Kaggle competition, of which there were 61578 images (taken from the GZ1 dataset) with their corresponding 37-class response vectors. The GZ1 data was split into training, validation and testing subsets.

A ResNet50 model, a CNN architecture, was trained using the GZ1 response vectors. The loss function used was the MSE and the model trained on the training set for 500 epochs (iterations). A set of self-defined rules, derived using the Galaxy Zoo decision tree in Figure 2.6, allowed a mapping from the 37-class vectors to a Hubble type. In certain instances, the response vectors didn't provide enough information to map the 37-class vectors to a specific spiral class, hence the labels 'Sx' and 'SBx' were used to represent non-barred and barred spirals, respectively. The flow diagram in Figure 2.7 illustrates the logic of the rules-based method. In Figure 2.7, the label 'Class1.1' indicates that the majority of participants chose the first response for question 1. When the response vectors didn't provided enough information to choose an exact spiral class, the classes used were 'Sx' and 'SBx'.

The 37-class vectors for the RSA catalogue samples were then predicted, using the ResNet50 model, and then the rules were used to map these vectors to a Hubble type. Transfer learning was also employed wherein the ResNet50 model trained using the GZ1 response vectors was used to initialise a model trained using the Hubble types in the RSA catalogue. Transfer Learning is employed because the RSA catalogue much smaller in size compared to the GZ1 data and the weights and parameters learned from the GZ1 will provide a strong foundation for learning to predict the Hubble types. To test Transfer Learning for galaxy classification, a ResNet50 is trained on the RSA catalogue without an initialisation and the performance is compared to that of the Transfer Learning model.

2.5 Crowd-sourced vs. expert data for galaxy classification methodology

The work described in this Section has been published in the proceedings of the *7th Intl. Conference on Soft Computing & Machine Intelligence* by (Variawa et al., 2020b).

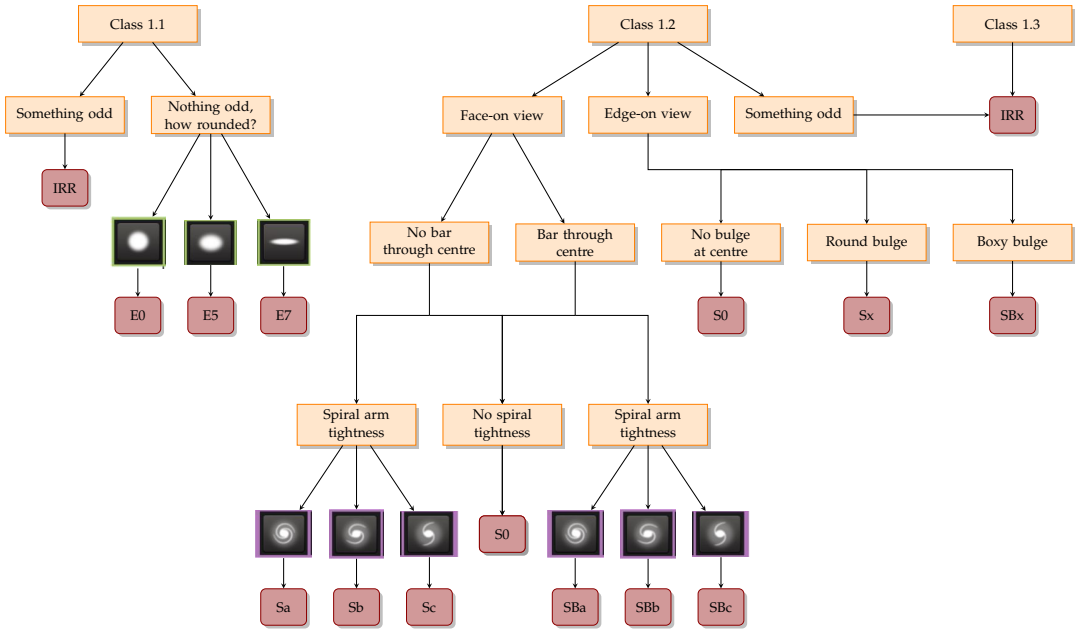

FIGURE 2.7: Rules-Based Logic To Map GZ1 Onto A Hubble Types

Given the popularity of combining machine learning with crowd-sourced labels for classifying galaxies (Hart et al., 2016), a set of experiments was designed to investigate the generalisation of these crowd-sourced labels to an expert classification.

A ResNet50 model was trained using the GZ2 response vectors. The data was split into training, validation and testing subsets. The SGD optimisation algorithm was used to train the model using 0.1 as the learning rate and 0.9 as the momentum value. Transfer Learning was then employed wherein this model was used to initialise two models; the first trained using the Hubble types provided in the GZ2-H data (Hart et al., 2016), and the second trained using the Hubble types in the EFIGI catalogue (Vaucouleurs et al., 2013). The Cross-Entropy loss function was used to train both Transfer Learning models. Both models trained using the Adam optimisation algorithm with 1×10^{-5} chosen as the learning rate and 5×10^{-5} chosen as the value for weight decay. The first model trained on the GZ2-H dataset for 178 epochs, where early-stopping on the loss on the validation subset determined the number of epochs. The second model trained on the EFIGI catalogue for 104 epochs, where early-stopping on the loss on the validation subset determined the number of epochs.

In the GZ2-H, the ellipticals were labelled as Er, Ei and Ec (Hart et al., 2016). Inspecting samples for these classes, they were identified as E0, E3 and E7, respectively. The GZ2-H only accounts for 3 elliptical classes (Hart et al., 2016) whereas the Hubble tuning fork has 4 and, as a result, the types E3 and E5 were grouped into one class, $E3-5$. The tuning fork in Figure 2.5 shows the classes used in the experiments, with E3 representing the joint $E3-5$ class.

Both the GZ2-H and EFIGI datasets were split into training, validation and testing

subsets using a stratified approach to ensure that every class was represented in each subset. The unseen EFIGI and GZ2-H test sets were then used to draw a comparison between the results from the models with the current state-of-the-art. To draw a comparison between the generalisation of the model trained on the crowd-sourced GZ2-H and the generalisation of the model trained on the expertly-labelled EFIGI catalogue, the RSA catalogue (Sandage and Tammann, 1981) was used as a holdout, expertly-labelled test set.

2.6 Deep Metric Learning methodology

The work described in this Section has been submitted for publication in the *Neural Computing & Applications journal* by (Variawa et al., Submitted 2021).

Although it was shown that the results obtained using the Cross-Entropy loss function were state-of-the-art (Variawa et al., 2020a), the work in (Boudiaf et al., 2020) was followed to improve on the results by combining the Label Smoothing technique with the Cross-Entropy loss.

Transfer Learning was employed wherein the ResNet50 model, which was trained on the GZ2 37-class response vectors, was retrained on both the EFIGI and GZ2-H datasets using the Cross-Entropy loss combined with Label Smoothing. The one-hot encoded class vectors were smoothed according to the same rules described in (Boudiaf et al., 2020). For the true class, i.e. 1, the class probability was set to

$$1 - \epsilon,$$

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

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

where $\epsilon = 0.1$ and K represents the number of classes which was 9 for the EFIGI catalogue and 11 for the GZ2-H dataset. The second-to-last layer of the ResNet50 had a size of 128, representing the size of the embedding space used to train the DML models. To determine the optimal size for the embedding space, various sizes were tested (512, 128 and 37), with 128 proving to be the most effective for both the EFIGI catalogue and GZ2-H dataset. Because GZ2-H has 11 Hubble classes (Hart et al., 2016), and the EFIGI has 9 (Vaucouleurs et al., 2013), the GZ2-H models are tested on both 9 and 11 Hubble classes for comparison.

The models that had been trained using the Label Smoothing Cross-Entropy loss were fine-tuned using DML techniques to investigate if using DML techniques could improve on the current state-of-the-art in galaxy classification. The embedding space size for the DML models was kept consistent at 128.

Both triplet and pair mining were used for training models on both the GZ2-H and EFIGI datasets and the same pairwise loss functions were used for both datasets. The pairwise loss functions used are the contrastive (Wang et al., 2006), triplet (Weifeng, 2018), Proxy-NCA (Goldberger et al., 2004), Proxy Anchor and Normalised Softmax (Boudiaf et al., 2020) losses.

2.7 Conclusion

This Chapter was opened by discussing the software and hardware used in the experiments. The Convolutional Neural Network architecture, which is a deep learning architecture that is employed in this research, was then discussed. Transfer learning and Deep Metric Learning, which are deep learning techniques that have been employed extensively in this research, were introduced in this Chapter. The Chapter then describes the crowd-sourced GZ1, GZ2 and GZ2-H datasets and the expert RSA and EFIGI catalogues. An analysis of each dataset is also provided. The Chapter then goes on to describe the setup for the experiments to test the effectiveness of machine learning for galaxy classification, compare the generalisation between crowd-sourced vs. expert labels for galaxy classification and, finally, investigate if a classification model can be fine-tuned using DML techniques in order to improve on the current state-of-the-art results in galaxy classification.

3 Results and Discussion

3.1 Introduction

Sections 3.2, 3.3 and 3.4 present the results from the experiments as well as provide comparisons with the current state-of-the-art in galaxy classification. Section 3.5 provides a discussion around the findings of the experiments.

3.2 Effectiveness of machine learning techniques for galaxy classification results

The results presented in this Section have been published in the proceedings of the *23rd International Conference on Information Fusion* by (Variawa et al., 2020a).

A ResNet50 model was trained using the GZ1 37-class response vectors. This model achieved a 0.0941 RMSE value on the unseen Kaggle test, whereas the competition's winning RMSE, obtained by (Dieleman et al., 2015), was 0.0749. The 0.0941 RMSE, however, manged to rank among the top 11% of entries into the competition (as of March 2020).

The 37-class vectors were predicted for the RSA catalogue samples, using the ResNet50 model. The set of defined rules, discussed in Section 2.4, then used these vectors to predict a Hubble type. In some instances, the rules were unable to choose a specific spiral type using the 37-class vectors. Specifically, some samples were described as having a bulge in the center and a bar through the center which may be classified as a barred spiral but also had an edge-on view which made it impossible to choose a specific barred spiral. As a solution to this, non-barred spirals were classified as Sx and barred spirals were classified as SBx. If the actual type of a sample was Sa, Sb or Sc and the predicted type was Sx, then the prediction was considered to be correct. If the actual type of a sample was SBa, SBb or SBc and the predicted type was SBx, then the prediction was considered to be correct.

Using this approach, an an accuracy of 20.69% was achieved when classifying the RSA samples with the rules. This result hinted at the potential for deriving the Hubble type of a galaxy using the GZ1 vectors.

Further investigating the predictions showed that using the rules-based approach, spirals were rarely predicted to be an elliptical and vice versa. This is an indication

of correlation between the GZ1 vectors and Hubble types. The result also indicates that classifying an exact spiral type from an image is challenging. This could be a result of low resolution images and also that the specific structures of both barred and non-barred spirals can look almost identical in images

Two ResNet50 models were trained on the RSA catalogue; the first had been initialised using the GZ1 dataset (i.e. instead of the model's weights being initially set to random values, the model was instead trained on the GZ1 dataset so that when the model is retrained on the RSA catalogue the weights have already been trained and the model will thus be able to better converge on the RSA catalogue) and the second had not been initialised. The same loss function, weighted Cross-Entropy, was used to train both models.

The ResNet50 model, that had not been initialised, trained using the RSA catalogue yielded an accuracy of 30%, however, the confusion matrix provided in Table 3.1 shows that the model predicted all samples as the spiral class, this could be because 69% of samples in the RSA catalogue belong to the spiral class, as shown in Table 2.4, hence the model learned a bias towards the spiral class. The ResNet50 model, that had been initialised using the GZ1 data, achieved an accuracy of 32.7%, and Table 3.2 shows that the Transfer Learning model was better at predicting the Hubble types than the ResNet50 with no initialisation.

The Transfer Learning model outperformed both the ResNet-50 model with no initialisation and the rules-based approach. Table 3.2 shows that although the Transfer Learning model was generally able to successfully separate the spiral structures from the elliptical structures, there were still cases of spirals being classified as elliptical's and vice versa, which is indicative that at certain angles, elliptical and spiral structures look very similar and can be difficult to distinguish between the two structures. A deeper analysis of the predictions showed that the model still struggled to separate the spiral structures from one another.

TABLE 3.1: Revised Shapley-Ames Actual Hubble Type vs Predicted Hubble Type - **ResNet50 With No Initialisation**- Actual Instances

		Predicted Type			
		Barred	**Elliptical**	**Irr**	**Spirals**
Actual Type	**Barred**	0	0	0	235
	Elliptical	0	0	0	500
	Irr	0	0	0	31
	Spirals	0	0	0	485

TABLE 3.2: Revised Shapley-Ames Actual Hubble Type vs Predicted Hubble Type - **GZ1 Transfer Learning Model**- Actual Instances

		Predicted Type			
		Barred	**Elliptical**	**Irr**	**Spirals**
Actual Type	**Barred**	52	59	46	78
	Elliptical	24	225	64	187
	Irr	1	9	18	3
	Spirals	48	296	34	107

3.3 Crowd-sourced vs. expert data for galaxy classification results

The results presented in this Section have been published in the proceedings of the *7th Intl. Conference on Soft Computing & Machine Intelligence* by (Variawa et al., 2020b).

The GZ2 response vectors were used to train a ResNet50 model. The was split data into training, validation and testing subsets. The training and validation subsets were used during the training of the model whereas the test set was kept hidden from the training of the model. An RMSE of 0.2236 was achieved on the unseen test set.

The model trained on the 37-class response vectors in the GZ2 dataset was used to initialise two models; the first was trained using the GZ2-H dataset and the second was trained using the EFIGI catalogue. Both the GZ2-H and EFIGI sets were split into training, validation and testing subsets using a stratified approach to ensure that every class was represented in each subset. The GZ2-H and EFIGI testing subsets were used to draw a comparison between the results of the ResNet50 models with the current state-of-the-art results. The expert RSA catalogue was then used to compare the generalisation between the crowd-sourced GZ2-H and expertly-labelled EFIGI datasets.

The ResNet50 model trained on the GZ2-H dataset achieved an accuracy of 85.22% on the 11-class test set, significantly outperforming the 65.20% reported by (Barchi et al., 2020). Testing on the RSA catalogue, using 11 classes, the accuracy was 25.7% and the unweighted F_1 score was 0.2035. Predicting 9 classes in the RSA catalogue, the accuracy was 30.50% and the unweighted F_1 score was 0.2383. These results indicate that although the ResNet50 models improved on the current state-of-the-art for the GZ2-H dataset, the crowd-sourced classifications were not able to generalise to the expert Hubble tuning fork.

The ResNet50 model that had been trained using the EFIGI catalogue achieved an accuracy of 97.76% for 3 classes, marginally outperforming the 97.27% reported by (Khalifa et al., 2017). When predicting 9 classes in the EFIGI catalogue, an accuracy of 82.72% was achieved. Predicting 9 classes in the RSA catalogue, the accuracy was 29.52% and the unweighted F_1 score was 0.2822.

In Table 3.3 a comparison is provided between then ResNet50 results on the EFIGI catalogue and the results presented by (Khalifa et al., 2017). The table also provides a comparison between the ResNet50 results on the GZ2-H 11-class prediction and the results presented by (Barchi et al., 2020). Table 3.3 illustrates that the ResNet50 models were able to improve on the state-of-the-art accuracy on both the EFIGI and GZ2-H datasets.

TABLE 3.3: Classification Accuracy For Proposed Techniques

Paper	Model Arch'	Model Init' on	Fine Tuning	Number Classes	Loss Function	Test set Accuracy (%)
(Khalifa et al., 2017)	CNN	EFIGI	EFIGI	*3*	CE	*97.27*
(Variawa et al., 2020a)	Res50	GZ2	EFIGI	*3*	CE	**97.76**
(Barchi et al., 2020)	CNN	GZ2-H	GZ2-H	*11*	CE	*65.20*
(Variawa et al., 2020a)	Res50	GZ2	GZ2-H	*11*	CE	**85.20**

In Figure 3.1, Receiver Operating Characteristic (ROC) curves are presented for the ResNet50 models trained on the EFIGI and GZ2-H data. Figure 3.1 shows that the GZ2-H 9-class curve is higher than the EFIGI 9-class curve which may indicate that the model trained on the GZ2-H data generalised better to the expert RSA catalogue. However, the model trained using the EFIGI catalogue had a higher F_1 score (0.2822) than the model trained using the GZ2-H dataset (0.2383). Table 3.5 presents a confusion matrix which shows that the model trained on the EFIGI catalogue was better at predicting the Hubble types Irr, SBa and Sa than the model trained on the GZ2-H dataset as shown in the confusion matrix in Table 3.4. It is also shown that neither model was able to generalise well to the RSA catalogue. This result was unable to conclude whether an expertly-labelled catalogue is better for automating galaxy classification using machine learning than a crowd-sourced dataset, necessitating further investigation.

TABLE 3.4: Revised Shapley-Ames Actual Hubble Type vs Predicted Hubble Type - **GZ2-H (9 Classes) ResNet50 Cross-Entropy Model**- MAP@1

		Predicted Type								
		E	**Irr**	**S0**	**SBa**	**SBb**	**SBc**	**Sa**	**Sb**	**Sc**
Actual Type	**E**	.64	.02	.14	.0	.01	.17	.0	.01	.01
	Irr	.22	.05	.27	.0	.02	.20	.0	.0	.24
	S0	.41	.01	.36	.0	.02	.14	.0	.03	.03
	SBa	.08	.05	.17	.0	.29	.15	.0	.09	.17
	SBb	.03	.0	.16	.0	.24	.35	.0	.05	.17
	SBc	.09	.0	.29	.0	.07	.39	.0	.03	.13
	Sa	.28	.01	.20	.0	.07	.17	.01	.17	.1
	Sb	.13	.0	.25	.0	.08	.15	.0	.21	.18
	Sc	.08	.004	.17	.0	.04	.21	.0	.11	.39

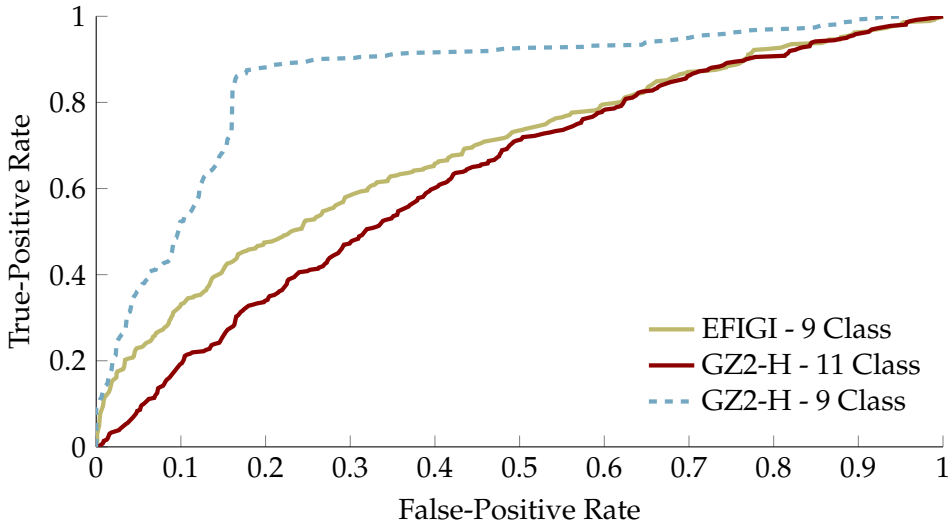

FIGURE 3.1: Receiver Operating Characteristic Curve For EFIGI And GZ2-H Classification Models Tested On The RSA Catalogue

TABLE 3.5: Revised Shapley-Ames Actual Hubble Type vs Predicted Hubble Type - **EFIGI (9 Classes) ResNet50 Cross-Entropy Model**- MAP@1

		Predicted Type								
		E	**Irr**	**S0**	**SBa**	**SBb**	**SBc**	**Sa**	**Sb**	**Sc**
Actual Type	**E**	.37	.0	.06	.0	.0	.0	.09	.38	.10
	Irr	.0	.17	.02	.02	.12	.09	.15	.24	.17
	S0	.23	.0	.20	.01	.01	.0	.17	.32	.07
	SBa	.0	.0	.25	.25	.14	.0	.05	.28	.05
	SBb	.0	.02	.04	.13	.28	.05	.07	.35	.05
	SBc	.0	.06	.0	.01	.26	.19	.06	.29	.13
	Sa	.10	.01	.20	.03	.05	.001	.22	.31	.05
	Sb	.02	.01	.04	.03	.10	.04	.15	.47	.14
	Sc	.0	.02	.03	.003	.15	.09	.07	.32	.31

3.4 Deep Metric Learning results

This Section presents the results which have been submitted for publication in the *Neural Computing & Applications journal* by (Variawa et al., Submitted 2021).

It has been shown that the ResNet50 architecture when combined with Transfer Learning is able to outperform the state-of-the-art accuracy for both the EFIGI and GZ2-H datasets (Variawa et al., 2020a). The results thus far have not shown a significant difference in the use of crowd-sourced vs. expert labels for galaxy classification, necessitating further investigation.

Section 2.5 described a ResNet50 model initialised on the 37-class vectors in the GZ2 data that had served as an initialisation for two models; one trained on the GZ2-H data and the other trained on the EFIGI catalogue, where both models were trained used the Cross-Entropy loss function. The ResNet50 model that had been

initialised on the GZ2 37-class vectors was retrained on both the EFIGI and GZ2-H datasets, using the Label Smoothing Cross-Entropy loss. The second-to-last layer of the ResNet50 models had a size of 128, representing the size of the embedding space, where 128 was shown to be the optimal size for both the EFIGI and GZ2-H datasets. The embedding size of 128 was kept consistent in the subsequent DML models.

Classification models trained using the Cross-Entropy loss function output the class probabilities for each sample and the predicted class is chosen as the one with the highest probability. For the ResNet50 models trained using Label Smoothing Cross-Entropy and the subsequent DML models, the output was instead chosen to be the 128-dim embedding vectors for both the train (EFIGI/GZ2-H) and test (RSA) sets. Using the 128-dim output, it is possible to study not only the accuracy of the models but also the information retrieval of the models. The Euclidean norm between the 128-dim test (RSA) vectors and the 128-dim train (EFIGI/GZ2-H) vectors was calculated. The predicted class for each test (RSA) sample was then chosen to be the class of the nearest neighbour in the train (EFIGI/GZ2-H) set (i.e. the class of the train sample which had the lowest Euclidean norm between the test sample). The accuracy was then calculated as the number of correct predictions divided by the size of the test (RSA) set. The MAP@R values were determined by calculating (for each test (RSA) sample) how many of the top R nearest neighbours (in the train (EFIGI/GZ2-H) samples) had the correct class.

The ResNet50 model trained on the EFIGI catalogue using the Label Smoothing Cross-Entropy loss function was used as the initialisation for the subsequent DML models trained on the EFIGI catalogue. Multiple models were trained on the EFIGI catalogue using the pairwise loss functions mentioned in Section 2.6. In order to reduce the effect of dataset dependence, 5-fold cross validation was used when training the models and the average MAP@R and accuracy values were calculated when testing on the RSA catalogue. The results showed that using the triplet mining function coupled with the Proxy-anchor loss function yielded the best mean average precision (MAP) value on the RSA catalogue at 0.3953 while the Normalised Softmax loss function yielded the best accuracy at 30.05%. The pair mining function coupled with the Proxy-NCA loss function yielded the best accuracy and MAP value on the RSA catalogue at 29.03% and 0.3917, respectively.

The ResNet50 model trained on the GZ2-H dataset using the Label Smoothing Cross-Entropy loss function was used as the initialisation for the subsequent DML models trained on the GZ2-H dataset. Multiple models were trained on the GZ2-H dataset using the pairwise loss functions mentioned in Section 2.6. In order to reduce the effect of dataset dependence, 5-fold cross validation was used when training the models and the average MAP@R and accuracy values were calculated when testing on the RSA catalogue. When testing on 11 Hubble classes in the RSA catalogue, the results showed that the triplet mining function coupled with the Proxy-NCA loss function yielded the highest accuracy at 24.45% and the best MAP value at 0.2955.

Again testing on 11 Hubble classes in the RSA catalogue, the pair mining function coupled with the Normalised Softmax loss function yielded the best accuracy and MAP values at 25.04% and 0.3102, respectively. Testing on 9 Hubble classes in the RSA catalogue, the results showed that using the triplet mining function coupled with the Normalised Softmax loss function yielded the best accuracy and MAP values at 30.88% and 0.3483, respectively. Again testing on 9 Hubble classes in the RSA catalogue, the results showed that the pair mining function coupled with the Proxy-NCA loss function yielded the best accuracy and MAP values at 30.32% and 0.3577, respectively.

Table 3.6 provides a consolidated view of the results obtained by the Label Smoothing Cross-Entropy models as well the improvements on both accuracy and MAP@R values achieved through the DML experiments.

TABLE 3.6: Deep Metric Learning Retrieval Results Showing The Datasets, Loss and Mining Munctions, Accuracy and MAP@R

Section	Model Init' on	Fine Tuning	Number Classes	Loss Type	Loss Function	Mining Function	Embedding Size	RSA Accuracy (%)	RSA MAP@R†
3.4	GZ2	EFIGI	*9*	Class	LS-CE	-	128	*23.50*	*0.3051*
	GZ2	GZ2-H	*9*	Class	LS-CE	-	128	*28.34*	*0.2145*
	GZ2	GZ2-H	*11*	Class	LS-CE	-	128	*25.30*	*0.1852*
	GZ2	EFIGI	*9*	Class	LS-CE	-	128	*24.86*	*0.3119*
	EFIGI (LS-CE)	EFIGI	*9*	DML	Contrastive	Pair	128	*27.97*	*0.3859*
	EFIGI (LS-CE)	EFIGI	*9*	DML	Norm-S'max	Triplet	128	**30.05**	*0.3889*
	EFIGI (LS-CE)	EFIGI	*9*	DML	P-Anchor	Triplet	128	*29.32*	**0.3953**
	EFIGI (LS-CE)	EFIGI	*9*	DML	P-NCA	Pair	128	**29.03**	**0.3917**
	EFIGI (LS-CE)	EFIGI	*9*	DML	Triplet	Triplet	128	*28.98*	*0.3898*
	GZ2	GZ2-H	*9*	Class	LS-CE	-	128	*24.32*	*0.3101*
	GZ2-H (LS-CE)	GZ2-H	*9*	DML	Contrastive	Pair	128	*26.84*	*0.3362*
	GZ2-H (LS-CE)	GZ2-H	*9*	DML	Norm-S'max	Triplet	128	**30.88**	**0.3483**
	GZ2-H (LS-CE)	GZ2-H	*9*	DML	P-Anchor	Triplet	128	*29.48*	*0.3350*
	GZ2-H (LS-CE)	GZ2-H	*9*	DML	P-NCA	Pair	128	**30.32**	**0.3577**
	GZ2-H (LS-CE)	GZ2-H	*9*	DML	Triplet	Triplet	128	*26.42*	*0.3328*
	GZ2	GZ2-H	*11*	Class	LS-CE	-	128	*20.85*	*0.2704*
	GZ2-H (LS-CE)	GZ2-H	*11*	DML	Contrastive	Pair	128	*23.86*	*0.3032*
	GZ2-H (LS-CE)	GZ2-H	*11*	DML	Norm-S'max	Pair	128	**25.04**	**0.3102**
	GZ2-H (LS-CE)	GZ2-H	*11*	DML	P-Anchor	Pair	128	*24.72*	*0.2909*
	GZ2-H (LS-CE)	GZ2-H	*11*	DML	P-NCA	Triplet	128	**24.45**	**0.2955**
	GZ2-H (LS-CE)	GZ2-H	*11*	DML	Triplet	Triplet	128	*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.

In order to provide a more in-depth investigation of the predictions, MAP@1 confusion matrices are provided, where MAP@1 is equivalent to the accuracy of each class:

- Table 3.7 shows the MAP@1 actual vs. predicted labels for the GZ2-H 11 class model.
- Table 3.8 shows the MAP@1 actual vs. predicted labels for the GZ2-H 9 class model.
- Table 3.9 shows the MAP@1 actual vs. predicted labels for the EFIGI 9 class model.

ROC curves in Figure 3.3 are provided to further analyse the true-positive vs false-positive rates for the 3 best DML models. The ROC curve for the EFIGI 9-Class predictions in Figure 3.3 has a spike in the True-Positive Rate which is indicative of the fact that at lower classification thresholds, the True-Positive Rate is higher which shows that the model predicts the correct classes with lower probabilities. Table 3.9 which shows that for the classes E, Irr and S0 An example image from the RSA catalogue, with Hubble type Sa, is shown in Figure 3.2 and its 11 nearest neighbours in the EFIGI catalogue.

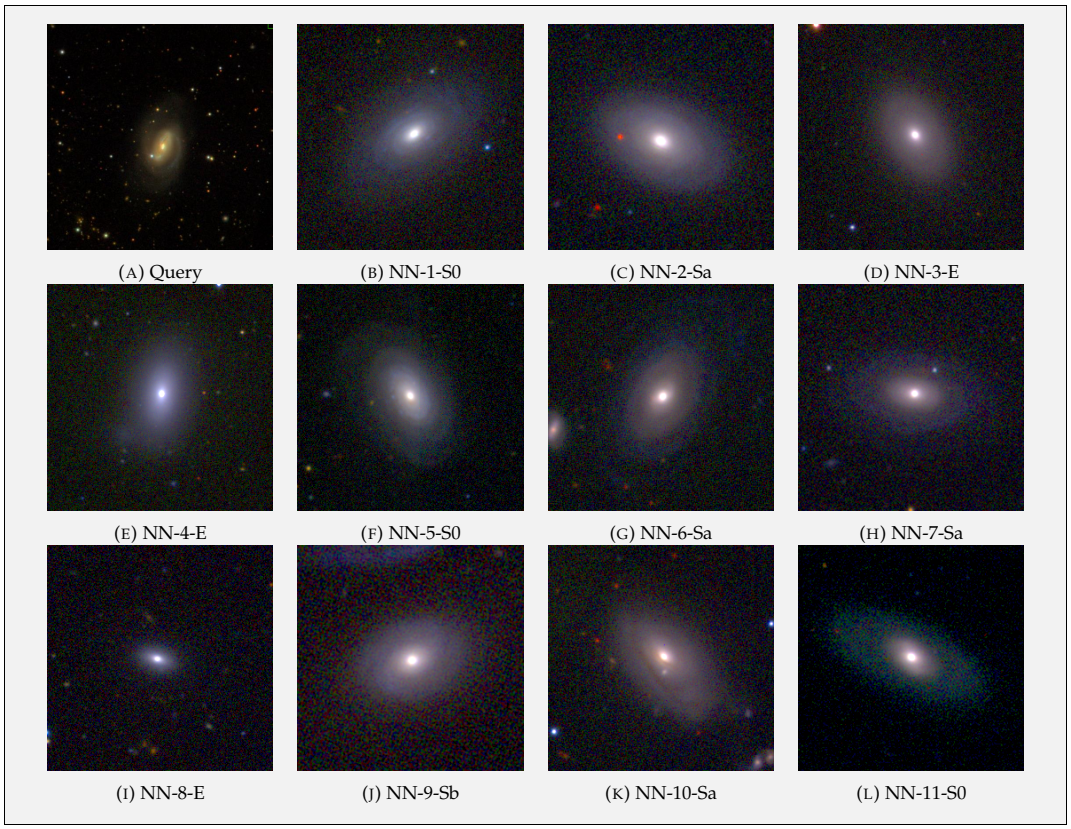

FIGURE 3.2: An Example Image From The RSA Catalogue, Hubble Type Sa, With The Top 11 Nearest Neighbours From The EFIGI Triplet/Proxy-NCA Model

Following the *DeepGalaxy* methodology described in (Cai et al., 2020), a ResNet50 model was trained on their simulated dataset. This model was used to generate the 1-D feature vectors for the EFIGI catalogue and a multi-layer perceptron (MLP) was trained on these 1-D vectors input and the one-hot encoded Hubble labels of the EFIGI as the output (targets). This model was then tested on the EFIGI unseen testing set as well as the RSA test set. On the EFIGI unseen testing set, the model achieved an accuracy of 28.8% compared to the 85.22% reported in Section 3.3. On the RSA test set, the accuracy achieved using their methodology was 18.33% compared to the 29.52% reported in Section 3.3.

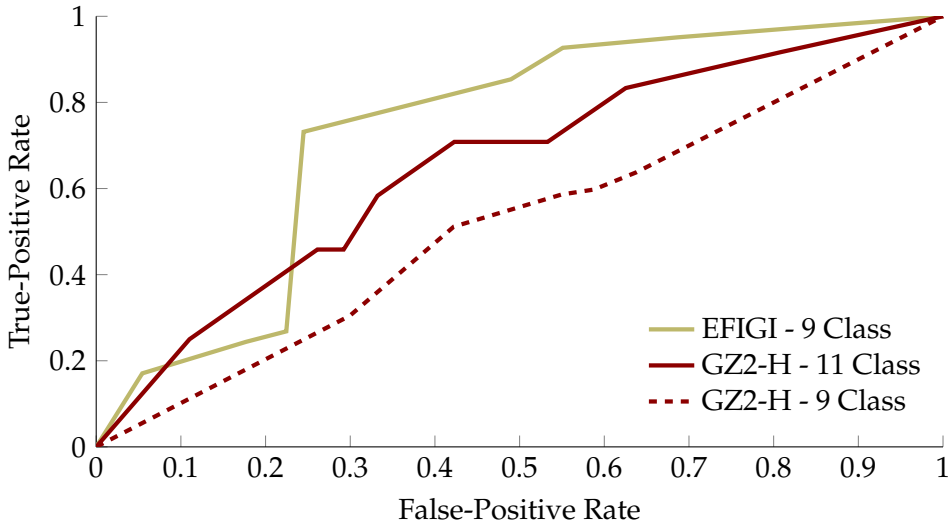

FIGURE 3.3: Receiver Operating Characteristic Curves For The EFIGI 9-Class Triplet/Proxy-NCA Model, GZ2-H 11-Class Triplet Proxy-NCA Model And GZ2-H 9-Class Pair/Proxy-NCA Model

3.5 Discussion

To investigate how effective machine learning is when applied to galaxy classification, a ResNet50 model was trained on the 37-class response vectors in the GZ1 dataset. A set of self-defined rules, derived the decision tree in Figure 2.6, allowed a mapping from the GZ1 37-class vectors to a Hubble type in the tuning fork in Figure 2.5. The logic for the set of rules is shown in Figure 2.7. The ResNet50 model was also used to initialise a model trained on the RSA catalogue to predict the Hubble types of galaxies. To test how effective Transfer Learning is when applied to galaxy classification, a ResNet50 was trained using the RSA catalogue with no initialisation. Using the RSA catalogue, the results shown that the Transfer Learning model achieved a better accuracy (32.7%) than the accuracy achieved using the set of rules (20.69%) and the accuracy achieved by the ResNet50 which had not been initialised (30%). This result indicated the effectiveness of using Transfer Learning with the ResNet50 architecture for galaxy classification.

Given that the effectiveness for galaxy classification when using Transfer Learning with the ResNet50 architecture has been shown, Transfer Learning with the ResNet50 was again employed to test the generalisation of crowd-sourced labels to the expert Hubble tuning fork. First, GZ2 vectors were used to train a ResNet50 model. This ResNet50 then served as the initialisation for two models; one trained using the crowd-sourced GZ2-H dataset and the other trained using the expertly-labelled EFIGI catalogue. The loss function used to train both models was Cross-Entropy. Testing on the RSA catalogue showed that although the ResNet50 models had outperformed the current state-of-the-art accuracy for the EFIGI and GZ2-H datasets, neither model could generalise well to the RSA catalogue. This result was insufficient to generally conclude whether an expertly-labelled catalogue is better

TABLE 3.7: Revised Shapley-Ames Actual Hubble Type vs Predicted Hubble Type - **Galaxy Zoo 2 (11 Classes) Triplet/Proxy-NCA model** - MAP@1

		Predicted Type									
		E0	**E3-5**	**E7**	**Irr**	**S0**	**SBa**	**SBb**	**SBc**	**Sa**	**Sb**	**Sc**
Actual Type	**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

TABLE 3.8: 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

for automating galaxy classification using machine learning than a crowd-sourced dataset, necessitating further investigation.

The ResNet50 model trained on the 37-class vectors in the GZ2 dataset was retrained on the EFIGI catalogue using the Label Smoothing Cross-Entropy loss. The model output the 128-dim vectors instead of the class probabilities, making it possible to study the information retrieval accuracy. The 128-dim vector for each sample in the RSA test set was compared to the 128-dim vector for each sample in the EFIGI train set using the Euclidean norm. The predicted class for each test sample was chosen to be the class of its nearest neighbour in the EFIGI train set. The accuracy obtained when testing on the RSA catalogue (9 classes) was 24.86% and the MAP@R value was 0.3119. The accuracy obtained by the Label Smoothing Cross-Entropy model, 24.86%, is lower than the accuracy achieved by the ResNet50 in Section 3.2, 30%, which is indicative of the difficult nature of learning the 128-dim embedding vectors

TABLE 3.9: Revised Shapley-Ames Actual Hubble Type vs Predicted Hubble Type - **EFIGI (9 Classes) Triplet/Proxy-NCA Model** - MAP@1

		Predicted Type								
		E	**Irr**	**S0**	**SBa**	**SBb**	**SBc**	**Sa**	**Sb**	**Sc**
Actual Type	**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

(i.e. the output of the Label Smoothing Cross-Entropy model) as opposed to learning the 9-class probability vectors (i.e. the output of the ResNet50 in Section 3.2. The 128-dim embedding vectors make it possible, however, to study not only the accuracy of the model but also the information retrieval accuracy (i.e. MAP@R value) of the model.

The ResNet50 model trained on the 37-class vectors in the GZ2 dataset was retrained on the GZ2-H dataset using the Label Smoothing Cross-Entropy loss. The model output the 128-dim vectors instead of the class probabilities, making it possible to study the information retrieval accuracy. The 128-dim vector for each sample in the RSA test set was compared to the 128-dim vector for each sample in the GZ2-H train set using the Euclidean norm. The predicted class for each test sample was chosen to be the class of its nearest neighbour in the GZ2-H train set. Using this approach, the accuracy obtained on the RSA catalogue (11 classes) was 20.85% and the MAP@R value was 0.2704. Testing on 9 classes in the RSA catalogue, the accuracy was 24.32% and the MAP@R value was 0.3101.

The Label Smoothing Cross-Entropy models showed that while the model trained on the GZ2-H achieved higher accuracy than the model trained on the EFIGI, the information retrieval (i.e. MAP@R) was a lot higher in the model trained on the EFIGI (i.e. better accuracy is not equivalent to better information retrieval).

Further, it was investigated whether DML methods can be used to fine-tune deep learning models that had been initialised using the Label Smoothing Cross-Entropy loss. The results from the experiments using DML techniques have allowed us to answer the second research question of this dissertation "Can a classification model be fine-tuned using Deep Metric Learning techniques to outperform state-of-the-art in galaxy classification?" where Table 3.6 shows that DML techniques were able to outperform the state-of-the-art in galaxy classification. The Proxy-NCA loss function has marginally outperformed the other pairwise loss functions on both the EFIGI

catalogue and GZ2-H dataset. Specifically, for classifying Hubble types in the RSA catalogue:

- Using the Label Smoothing Cross-Entropy loss function to train a ResNet50 model and then fine-tuning that model using the pair mining function and Proxy-NCA loss function produced the best overall 9-class GZ2-H model, yielding an accuracy of 30.32%, an improvement of 1.98%, and a MAP value of 0.3577, an improvement of 0.1432.
- Using the Label Smoothing Cross-Entropy loss function to train a ResNet50 model and then fine-tuning that model using the pair mining function and Proxy-NCA loss function produced the best overall EFIGI model, yielding an accuracy of 29.03%, an improvement of 5.53%, and a MAP value of 0.3917, an improvement of 0.0866.

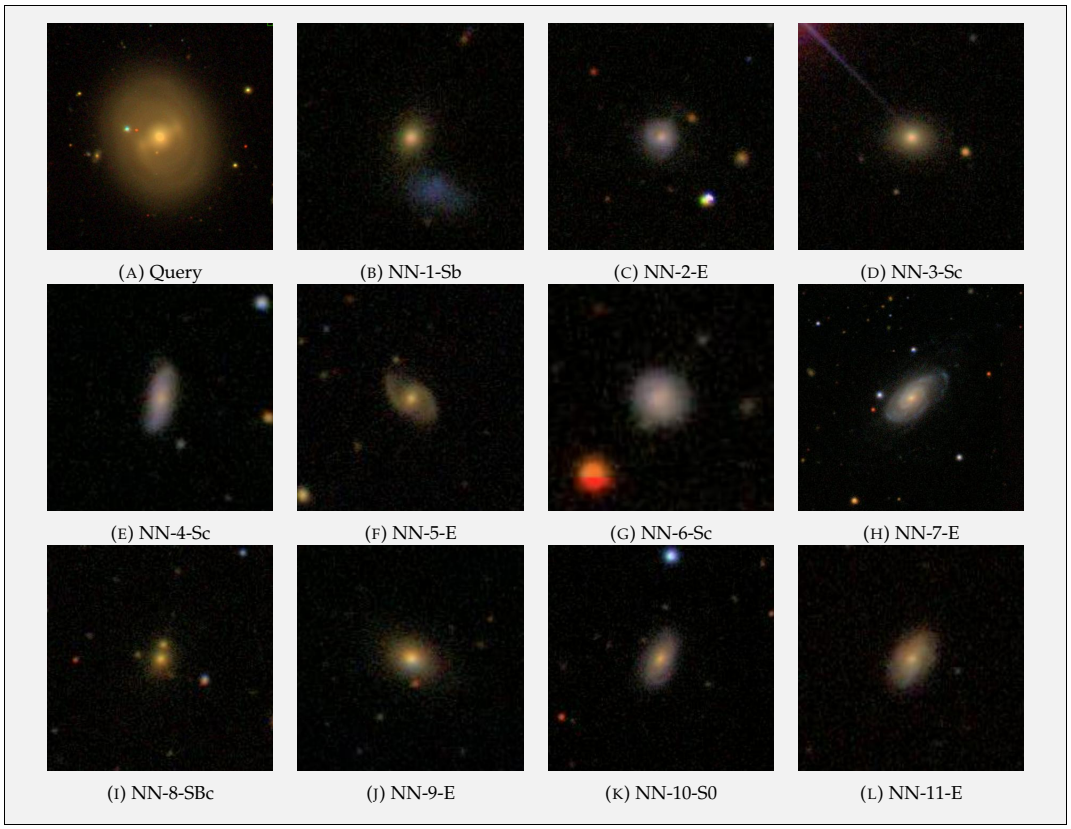

FIGURE 3.4: An Example Image From The RSA Catalogue, Hubble Type SBa, With The Top 11 Nearest Neighbours From The GZ2-H Pair/Normalised-Softmax Model

Through extensive experimentation, it has also been shown that crowd-sourced data seem on the surface to be effective for galaxy classification. The models trained on the crowd-sourced GZ2-H dataset produced accuracy and MAP values comparable to the models trained on the expertly labelled EFIGI catalogue, as shown in Table 3.6. However, studying the confusion matrices in Tables 3.7 and 3.8 shows that the models trained on the GZ2-H dataset were not able to predict the Irr, SBa and Sa classes.

This is due to these 3 classes being severely under-represented in the GZ2-H dataset, as shown in Table 2.4, relative to the other classes.

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 3.8. The misclassification of the Hubble type Sa as Sb and Sc indicates that the Sb and Sc classes directly follow the Sa class in the tuning fork as seen in Figure 2.5. Figure 3.4 provides an example image from the RSA catalogue, with Hubble type SBa (another under-represented class in the GZ2-H dataset) and its predicted nearest neighbours from the GZ2-H dataset. Figure 3.4 illustrates that the ResNet50 model, trained using the pair mining function with the Proxy-NCA loss function, can pick out sample images as the nearest neighbours with similar structures to the SBa sample even if it could not explicitly predict SBa. Table 3.9 shows that this problem is not prevalent in the models trained on the EFIGI catalogue, which is an indication that using an expertly labelled catalogue of galaxies is better for automating galaxy classification using machine learning than using a crowd-sourced dataset. This results provides an answer to the first research question of this dissertation "How well does a model trained using crowd-sourced labels generalise to the expert Hubble tuning fork classification scheme?"

Studying the ROC curves in Figure 3.3 is another indication (substantiating the claim above) that using a catalogue of galaxies with expert labels is better than using a dataset with crowd-sourced labels for using machine learning to automate galaxy classification. As seen in Figure 3.3, the ROC curve for the EFIGI 9 class rises above both the GZ2-H 11 class and 9 class curves. These ROC curves illustrate that the Triplet/Proxy-Anchor model trained using the EFIGI dataset was better at predicting the actual Hubble types of galaxies in the RSA catalogue than both the Triplet/Proxy-NCA model trained on the GZ2-H 11-class data and the Pair/Proxy-NCA model trained on the GZ2-H 9-class data.

Although the *DeepGalaxy* framework described by (Cai et al., 2020) 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), the results have shown that using their framework did not transfer well to the task of classifying the Hubble types of galaxies.

3.6 Conclusion

This Chapter presented the results from the experiments as well as provided a comparison of the results with the current state-of-the-art in galaxy classification. An in-depth discussion of the findings has also been provided. The results have shown the effectiveness of machine learning techniques, specifically Transfer Learning, over the use of self-defined rules for galaxy classification. Studying classification models

trained using the Cross-Entropy loss function, has been insufficient to state specifically whether crowd-sourced or expert data is better for automating galaxy classification using machine learning. Through the use of classification models trained using the Label Smoothing Cross-Entropy loss (which output 128-dim embedding vectors instead of class probabilities) the results have shown that higher accuracy is not equivalent to better information retrieval. It has also been shown that classification models can be fine-tuned using DML techniques to improve on the current state-of-the-art results in galaxy classification. Finally, through extensive investigation, the results have shown that the expertly-labelled EFIGI catalogue is better for automating galaxy classification using machine learning than the crowd-sourced GZ2-H dataset.

4 Conclusion

The type and formation of galaxies often offer clues and insights into the beginning and development of the universe, making galaxy classification one of the important areas of study. Owing to the increased availability of images of galaxies, it has become increasingly popular to train machine learning models on crowd-sourced data to "automate" classification. Given the popularity of crowd-sourced labels for data, it is imperative 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.

A set of rules mapping the GZ1 response vectors to Hubble types in the Hubble tuning fork was defined. A ResNet50 model was trained using the GZ1 37-class vectors and then retrained using the Hubble types in the RSA catalogue. A second ResNet50 model was trained using the RSA catalogue with no initialisation. The results from the experiments showed that the Transfer Learning model achieved a higher accuracy than the set of rules and a higher accuracy than the model trained with no initialisation, showing the effectiveness of machine learning techniques (specifically Transfer Learning) in galaxy classification.

A ResNet50 model was then trained using GZ2 response vectors and Transfer Learning was employed wherein that model was used to initialise two models; the first trained using the Hubble types in the crowd-sourced GZ2-H dataset and the other trained using the Hubble types in the expert-labelled EFIGI catalogue. To draw a comparison between the generalisation of the model trained on crowd-sourced data and the model trained on expert data, the RSA catalogue served as an expert, hold-out test set. The results showed that the model trained using the expert EFIGI catalogue marginally outperformed the model trained on the crowd-sourced GZ2-H dataset. The difference, however, was not significant and necessitated further investigation.

The ResNet50 model trained on the 37-class vectors in the GZ2 dataset was used as an initialisation for two models; one trained using the Hubble types in the crowd-sourced GZ2-H dataset and the other trained using the Hubble types in the expertly-labelled EFIGI catalogue. Both these Transfer Learning models were trained using the Label Smoothing Cross-Entropy loss. Instead of outputting class probabilities, the Label Smoothing Cross-Entropy models output 128-dim embedding vectors, making it possible to study the information retrieval accuracies of the models.

The results showed that the ResNet50 model trained on the crowd-sourced GZ2-H dataset yielded a higher accuracy but the ResNet50 model trained using the expert EFIGI catalogue yielded a higher MAP value (i.e. better information retrieval). This result is indicative that higher accuracy is not equivalent to better information retrieval.

DML methods were then used to fine-tune the models that had been initialised using Label Smoothing Cross-Entropy. The size of the embedding space was kept consistent at 128 in the DML models. Both triplet and pair mining were used with various pairwise loss functions. Through the use of confusion matrices and ROC curves, the results 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. 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 3.7 and 3.8, because these classes are severely underrepresented in the GZ2-H data, as shown in Table 2.4, a problem which is not present in the EFIGI and RSA data. The results have also shown that a model trained using Label Smoothing Cross-Entropy can be fine-tuned using DML techniques to outperform the state-of-the-art in galaxy classification.

These results lead to four significant findings. First, machine learning techniques are effective in galaxy classification. Second, crowd-sourced labels should be used with caution. Third, the use of DML is a viable method of improving on the current state-of-the-art results in galaxy classification and related tasks. Fourth, Transfer Learning from crowd-sourced labelled data to expert-labelled data leads to significant improvement in classification 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.

Bibliography

[1] P.H. Barchi et al. "Machine and Deep Learning applied to galaxy morphology-A comparative study". In: *Astronomy and Computing* 30 (2020), p. 100334.

[2] Malik Boudiaf et al. "A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses". In: *arXiv:2003.08983v2* (July 2020).

[3] Maxwell X. Cai et al. "DeepGalaxy: Deducing the Properties of Galaxy Mergers from Images Using Deep Neural Networks". In: *arXiv:2010.11630v1* (Oct. 2020).

[4] Jorge de la Calleja and Olac Fuentes. "Machine learning and image analysis for morphological galaxy classification". In: *Monthly Notices of the Royal Astronomical Society* 349.1 (Apr. 2003), pp. 87–93.

[5] Jia-Ming Dai and Jizhou Tong. "Galaxy Morphology Classification with Deep Convolutional Neural Networks". In: *Astrophysics and Space Science* 364.4 (July 2018).

[6] Sander Dieleman, Kyle W. Willett, and Joni Dambre. "Rotation-invariant convolutional neural networks for galaxy morphology prediction". In: *Monthly Notices of the Royal Astronomical Society 450.2* (Mar. 2015).

[7] Zsolt Frei et al. "A Catalog of Digital Images of 113 Nearby Galaxies". In: *arXiv:astro-ph/9510062v1* (Oct. 1995).

[8] Alexandre Gauthier, Archa Jain, and Emil Noordehz. "Galaxy Morphology Classification". In: (Dec. 2016).

[9] Jacob Goldberger et al. "Neighbourhood Components Analysis". In: *Advances in Neural Information Processing Systems 17 (NIPS)* (2004).

[10] Raia Hadsell et al. "Dimensionality Reduction by Learning an Invariant Mapping". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2006).

[11] Ross E. Hart et al. "Galaxy Zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias". In: *Monthly Notices of the Royal Astronomical Society* 461.4 (2016), pp. 3663–3682.

[12] Swati Hira, Anita Bai, and Sanchit Hira. "An automatic approach based on CNN architecture to detect Covid-19 disease from chest X-ray images". In: *Applied Intelligence* (Nov. 2020).

[13] Siddhartha Kasivajhula, Naren Raghavan, and Hemal Shah. "Morphological Galaxy Classification Using Machine Learning". In: (Apr. 2003).

[14] Reza Katebi et al. "Galaxy morphology prediction using capsule networks". In: *arXiv:1809.08377* (Sept. 2018).

[15] Nour Eldeen M. Khalifa et al. "Deep galaxy: Classification of galaxies based on deep convolutional neural networks". In: *arXiv preprint arXiv:1709.02245* (2017).

[16] Jonathan Krause et al. "3D Object Representations for Fine-Grained Categorization". In: *4th IEEE Workshop on 3D Representation and Recognition, at ICCV 2013 (3dRR-13). Sydney, Australia* (Dec. 2004).

[17] Chris Lintott et al. "Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies". In: *Monthly Notices of the Royal Astronomical Society* 410.1 (2011), pp. 166–178.

[18] Vesna Lukic and Marcus Bruggen. "Galaxy Classification with Deep Learning". In: *Proceedings of the International Astronomical Union 12.S325* (2016).

[19] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering 22.10* (Oct. 2010).

[20] Allan Sandage and G.A. Tammann. "A Revised Shapley-Ames Catalog of Bright Galaxies". In: *Carnegie Institution Of Washington Publication 635* (1981).

[21] Mohamed Z. Variawa, Terence L. van Zyl, and Matthew Woolway. "A rules-based and Transfer Learning approach for deriving the Hubble type of a galaxy from the Galaxy Zoo data". In: *23rd International Conference on Information Fusion (FUSION 2020)* (2020). DOI: IEEEcatalognumber : CFP18FUS - ART , ISBN : 978-0-964527-6-2.

[22] Mohamed Z. Variawa, Terence L. van Zyl, and Matthew Woolway. "Comparing Generalisation Using Crowd-sourced vs Expert Labels for Galaxies Classification". In: *7th Intl. Conference on Soft Computing & Machine Intelligence (ISCMI 2020)* (2020).

[23] Mohamed Z. Variawa, Terence L. van Zyl, and Matthew Woolway. "Transfer Learning and Deep Metric Learning for Automated Galaxy Morphology Classification". In: *Neural Computing & Applications* (Submitted 2021).

[24] Mohamed Z Variawa, Terence van Zyl, and Matthew Woolway. *Galaxy classification*. https://github.com/mohamedzayyan/galaxyclassification. 2020.

[25] Gérard de Vaucouleurs et al. *Third Reference Catalogue of Bright Galaxies: Volume III*. Vol. 3. Springer Science & Business Media, 2013.

[26] C. Wah et al. *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. rep. CNS-TR-2011-001. California Institute of Technology, 2011.

[27] Mike Walmsley et al. "Galaxy Zoo: Probabilistic Morphology through Bayesian CNNs and Active Learning". In: *arXiv:1905.07424* (May 2019).

[28] Xun Wang et al. "Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning". In: *Proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2006).

[29] Ge Weifeng. "Deep Metric Learning with Hierarchical Triplet Loss". In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[30] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. "A survey of transfer learning". In: *Journal of Big Data* 3.9 (2016).

[31] B. Wu. "An introduction to neural networks and their applications in manufacturing". In: *Journal of Intelligent Manufacturing* 3 (Dec. 1992), 391–403.

[32] Rikiya Yamashita et al. "Convolutional neural networks: an overview and application in radiology". In: *Insights Imaging* 9 (June 2018), 611–629.