Denoising adversarial autoencoders: classifying skin lesions using limited labelled training data

Antonia Creswell1, Alison Pouplin1,2, Anil A. Bharath1,3
1BICV, Bioengineering, Imperial College, London, UK
2DAMAE Medical, Paris, France
3Data Science Institute, London, UK
E-mail: ac2211@ic.ac.uk

Abstract: The authors propose a novel deep learning model for classifying medical images in the setting where there is a large amount of unlabelled medical data available, but the amount of labelled data is limited. They consider the specific case of classifying skin lesions as either benign or malignant. In this setting, the authors’ proposed approach – the semi-supervised, denoising adversarial autoencoder – is able to utilise vast amounts of unlabelled data to learn a representation for skin lesions, and small amounts of labelled data to assign class labels based on the learned representation. They perform an ablation study to analyse the contributions of both the adversarial and denoising components and compare their work with state-of-the-art results. They find that their model yields superior classification performance, especially when evaluating their model at high sensitivity values.

1 Introduction

The problem of image classification is one of assigning one or more labels to a given image. Deep learning has been demonstrated to be able to achieve both human and super-human levels of performance [1] on classification tasks. However, achieving competitive levels of performance using deep learning often requires vast numbers of \{image, label\} pairs, typically in the millions.

In the medical image setting, it is unlikely that vast amounts of labelled images are available, particularly since medical expertise is required to label the data, which may be very costly and time consuming. Instead, it is often the case that there exists a large corpus of unlabelled data and a smaller set of labelled data.

We propose a model that is able to learn both from labelled data and from unlabelled data by building on previous work involving autoencoders [2–6]. Autoencoders are able to learn data representations from unlabelled data, by jointly learning an encoder and decoder. The encoder maps data samples – in this case images – to a low-dimensional encoding space, and the decoder maps the encoding back to image space. An autoencoder is trained to reconstruct its input. There are two key factors that enhance the performance of autoencoders, these are

• **Denoising**: Before encoding an input image, the image is corrupted, and the decoder is trained to recover the clean image. By making the decoding process more challenging, the autoencoder may learn more robust representations [6, 7].
• **Regularisation**: Rather than allowing encoded images to occupy an unconstrained space, the distribution of encoded images may be shaped to match a desired, prior distribution, for example a multivariate standard normal distribution. Regularisation reduces the amount of information that may be held in the encoding, forcing the model to learn an efficient representation of the image data.

To implement denoising in an autoencoder, an arbitrary corruption process may be used. For instance, white Gaussian noise [2] may be added to images sampled from the training data. Indeed, corruption is often trivial to implement. More challenging is the regularisation of the distribution of encoded images. There are at least two approaches for shaping the distribution of encoded images to match the desired distribution. The two key methods for regularising the encoding space are

**Variational**: Minimising the KL divergence between the distribution of encoded images and a chosen prior distribution [4]. For ease of implementation, the prior distribution is often a multivariate standard normal distribution and the encoder is designed to learn parameters of a Gaussian distribution.

**Adversarial**: Rather than using the encoder to parametrise a distribution and calculate the KL divergence, a third, discriminative model is trained to correctly distinguish encoded images from encodings drawn from a chosen prior distribution. The encoder is then updated to encode images such that the discriminator cannot distinguish between encoded images and encodings drawn from the prior distribution [5]. We will more formally introduce adversarial training in Section 2.2.

The adversarial [5] approach allows the encoder to be more expressive than the variational approach [4] and has achieved superior classification performance in a semi-supervised fashion on several benchmark dataset. While denoising and adversarial training have been used to augment autoencoders in isolation, they have yet to be combined in one model. Here, we propose augmenting an autoencoder with both a denoising criterion and by using adversarial training to shape the distribution of encoded image samples. We augment this model further to make use of labelled image data, where it is available, while still learning from unlabelled image data where label information is not available.

Our contributions are as follows:

• We introduce the semi-supervised denoising adversarial autoencoder (ssDAAE) which is able to learn from a combination of labelled and unlabelled image data (Section 2.3).
• We apply our model, the ssDAAE, to the task of classifying skin-lesions as benign or malignant in the setting where the amount of labelled image data is limited (Section 3).
• We compare the performance of the ssDAAE with a semi-supervised adversarial autoencoder (ssAAE), a fully supervised AAE (6AAE), and a fully supervised denoising adversarial autoencoder (sDAAE). We also compare the performance with a CNN having the same architecture as the encoder of the ssAAE.
and ssDAAE. Additionally, we assessed the effect of additive noise during training of the otherwise standard CNN. Our results show that the ssDAAE outperforms the others models on average for a high sensitivity.

- Generally, it is important to consider the reproducibility of results when applying deep learning techniques to medical data. In line with this, we perform a reproducibility experiment (Section 5) and explain several architectural changes necessary to achieve more stable training and hence less varied results between models that are trained from scratch with similar hyper-parameters.

Previous research has been conducted on the International Skin Imaging Collaboration (ISIC) [8] dataset, as part of a challenge hosted by the International Symposium on Biomedical Imaging (ISBI) [9]. The models that performed best during the challenge used fully-supervised deep learning architectures and transfer learning. In Table 1, we compare our results to those obtained by the top three participants, who each used pre-trained models (VGG-16 [10], ResNets [11, 12]) – specifically, this means that the network weights were pre-trained on a large source of natural image scenes. Further, participants either used additional sources of skin lesion images [10] or combined additional segmentation annotation with data augmentation [11]. In contrast, our results are obtained using only the ISIC dataset, with no pre-training.

Participants in the ISIC challenge reported results for three categories of skin lesion, however, we consider the critical problem of binary classification – melanoma-versus-all – in this study. Although we demonstrate this approach on skin lesions, the semi-supervised approach explored in this paper is not modality-specific, and could potentially be applied to other image datasets where labelled samples are in limited supply, but there is a surplus of unlabelled images that have been captured.

2 Method: classifying skin lesions

In this section, we formulate the ssDAAE. First, we discuss the skin lesion classification problem. Second, we describe the Adversarial Autoencoder (AAE) and then we describe how the AAE may be augmented to become the ssDAAE. Finally, we describe how the ssDAAE is trained.

2.1 Skin lesion classification

Skin lesion classification is a non-trivial problem. Humans have to be trained to be able to distinguish between benign (not harmful) skin lesions from malignant (harmful) skin lesions. Examples of benign and malignant skin lesions are shown in Fig. 1. The high-level goal is to train a model to correctly predict whether a skin lesion is benign or malignant. Beyond this, we want to design models for which we can be confident that we correctly identify a specific proportion of malignant skin lesions as being malignant, while still being able to correctly identify a large number of benign skin lesions as being so. To this end, in the following sections, we describe the model that we propose for skin lesion classification in the setting of limited labelled data.

2.2 Adversarial autoencoders

An autoencoder consists of two models, an encoder and a decoder, each with their own set of learnable parameters. In our approach, we used deep convolutional neural networks (CNNs) to embody the encoder and decoder. The encoder, \( E_{\theta_E} : x \rightarrow \hat{z} \) with parameters \( \theta_E \), is designed to map an image sample, \( x \) to an encoding, \( \hat{z} \). The encoded vector, \( \hat{z} \), is of much lower dimension than the number of pixels in an image, \( x \). The decoder \( D_{\theta_D} : \hat{z} \rightarrow \tilde{x} \) is designed to map an encoding \( \hat{z} \) back to an image, \( \tilde{x} \). The parameters, \( \theta_E \) and \( \theta_D \) of the encoder and decoder, respectively, are learned such that the difference between the input to the encoder, \( x \), and the output of the decoder, \( \tilde{x} \), are minimised.

The AAE [5] incorporates adversarial training [13] to shape the distribution of encoded data samples to match some chosen prior distribution, \( p(z) \), such as a multivariate standard normal distribution. Note that we are applying adversarial training to the encoded data samples, rather than the data samples, as more commonly seen in the literature [13, 14]. Adversarial training requires the introduction of another model, a discriminator, for which we also use a deep CNN. The discriminator, \( T_{\theta_T} : z \rightarrow (0, 1) \) maps encodings (either encoded data samples \( \hat{z} \) or samples drawn from the prior, \( z \)) to a probability of whether that sample comes from the chosen prior distribution. The parameters, \( \theta_T \) of the discriminator, are learned such that high values are assigned to samples that come from the chosen prior distribution and low values are assigned to samples that come from the encoder. To encourage encoded samples to match the chosen prior distribution, the parameters of the encoder, \( \theta_E \), are updated such that \( T_{\theta_T}(E_{\theta_E}(x)) \) are maximised.

Formally, the following objectives must be optimised during the training of an AAE:

\[
\begin{align*}
\arg\min_{\theta_E, \theta_D} & \quad \| x - D_{\theta_D}(E_{\theta_E}(x)) \|^2_2 \\
\arg\max_{\theta_T} & \quad E_{z \sim p(z)} \log T_{\theta_T}(z) + \mathbb{E}_{\hat{z} \sim \hat{p}(\hat{z})} \log (1 - T_{\theta_T}(\hat{z}))
\end{align*}
\] (1) (2)

Table 1 Comparison of our classifiers (ssDAAE) with the top three participants [10–12] of the ISBI challenge, and the best fusion strategy [9]

| Method            | Dataset | SP82 | SP98 | SP95 | SP99 | AUC |
|-------------------|---------|------|------|------|------|-----|
| ssDAAE (ours)     | 5000 (+7000) | 0.82 | 0.79 | 0.75 | 0.68 | 0.89 |
| VGG-16 [10]       | 7544    | 0.75 | 0.59 | 0.39 | —    | 0.87 |
| ResNet [11]       | 2000    | 0.73 | 0.55 | 0.40 | —    | 0.86 |
| ResNet [12]       | 3500    | 0.73 | 0.59 | 0.37 | —    | 0.87 |
| L-SVM [9]         | 2000    | 0.83 | 0.69 | 0.57 | —    | 0.89 |

*Those results are available in the ISBI paper [9]. SPX represents the specificity measured at x% sensitivity. AUC represents the area under the ROC curve. Numbers in bold indicates the best result for each column.*

Fig. 1 Examples of benign and malignant skin lesions. Classifying skin lesions as benign or malignant is non-trivial and requires expert knowledge.

(a) Benign, (b) Malignant
where \( p(x) \) is some chosen prior distribution, for example a standard normal, and \( D(x) \) is the training data distribution.

Equation (1) is the reconstruction cost that is used to train the encoder and decoder. This cost should be minimised, so that input images may be recovered after encoding and decoding. Equation (2) is the discriminator cost, which, when maximised, means that the discriminator can correctly distinguish between encoded data samples and samples from the chosen prior distribution. Equation (3) is a second regularisation cost used to update the encoder. When maximised – simultaneously with (3) [13] – this regularisation cost encourages the distribution of encoded samples to be similar to the chosen prior distribution.

An AAE may be trained entirely with unlabelled data and may be evaluated by measuring reconstruction error on a test dataset, or synthesising novel samples by first drawing samples, \( z \), from the chosen prior distribution and passing these through the decoder to produce synthetic images, \( \hat{x} \). The process of encoding and decoding test samples often reveals whether or not the decoder model has learned a sufficient representation for the data. A further test is to attempt to generate novel samples, by passing random encodings – drawn from the chosen prior distribution – through the decoder. Since a regularised autoencoder is able to generate novel samples, we often refer to it as a generative model of the data.

In its current form, it is not immediately obvious how an AAE may be used to perform classification. In fact, it is necessary to augment the encoder to predict not only the encoding but also the label.

### 2.3 Semi-supervised denoising AAE

Before learning to classify skin lesion as benign or malignant, we may first consider learning more about what a skin lesion looks like. This could involve learning the colour, general shape and textures of skin lesions. A ssDAAE allows us to do this by incorporating both a generative and classification model in one. The ssDAAE differs from the AAE in two ways.

First, the AAE is augmented by applying a corruption process. The corruption process \( C: x \rightarrow \tilde{x} \) is a stochastic process in which Gaussian noise, with standard deviation, \( \sigma \), is added to a data sample, \( x \) to obtain, \( \tilde{x} \). This change results in a denoising AAE (DAAE).

Second, the encoder of the DAAE is altered to define an ssDAAE by splitting the encoder into three parts. An initial encoder, \( E_{\theta_E}: x \rightarrow \tilde{h} \), and two sub-encoders, \( E_{\theta_E}': \tilde{h} \rightarrow \tilde{y} \) and \( E_{\theta_E}'': \tilde{h} \rightarrow \tilde{z} \). The encoder is trained to predict not only an encoding \( \tilde{z} \) but also a label vector, \( \tilde{y} \) (\( \tilde{y} \in \{0, 1\} \)). Adversarial training is used (as in an AAE [5]) to shape both the distribution of encoded samples to match a chosen prior distribution and the distribution of predicted class labels to match a categorical distribution [5].

Since we are posing skin lesion classification as a binary classification problem, we represent the labels benign and malignant using a single unit and apply a sigmoid function at the end of \( E_{\theta_E}' \). We can, therefore, train a label regulariser, \( T_{\theta_T} \), to distinguish predicted labels \( \tilde{y} \) from labels drawn from a binary distribution. This regularisation loss was intended to encourage the output of the classifier, \( E_{\theta_E}'' \), to be either 0 or 1 rather than taking values in between, and the first experiments conducted with this regulariser achieved good results (Section 4). However, while improving the model robustness, we noted that the adversarial nature of the label regularisation loss made the training less stable with small benefits. For the reproducibility experiments, we discarded the binary label discriminator (Section 5); this is a fair choice, given that such a regulariser process is not adopted in comparable learning methods/networks.

For an input \( x \), the output of the decoder \( \hat{x} \) is given by

\[
\hat{x} = D_{\theta_D}(E_{\theta_E}(C(x)))
\]

where \( [a, b] \) is a concatenation of \( a \) and \( b \).

The weights of the encoder, \( \theta_E \) and \( \theta_T \), are updated via both adversarial training to match the distribution of \( z \)'s to a chosen prior and to minimise a reconstruction error between \( x \) and \( \hat{x} \). This forms the generative part of the model and may be trained on entirely unlabelled data. This property of the model means that we can learn parameters \( \theta_E \) and \( \theta_T \) using large amounts of unlabelled data to learn more about the structure of skin lesions. We can also visualise what this model has learned by generating reconstructions of skin lesions and evaluating them by eye to see whether the model has captured the basic concept of what a skin lesion is.

Following on from this, we may use the limited labelled training data to ‘fine tune’ the generative model. We may use the labelled data to update the weights \( \theta_E \) or additionally to update \( \theta_T \) by minimising the classification error between predicted a label \( \tilde{y} \) and the true label \( y \). Experimentally, we found it beneficial to update both \( \theta_E \) and \( \theta_T \) as this made training more stable.

For completeness, note that – similar to an AAE – the weights of the decoder, \( \theta_D \), are learned as part of the minimisation of the reconstruction error between \( x \) and \( \hat{x} \). In Fig. 2, we present a diagram of our proposed model.

### 2.4 Training data

As described above, our ssDAAE may be trained using a mixture of both labelled and unlabelled data. The labelled data is obtained from the ISIC archive [8]. The archive consists of nearly 14k images, of which over 9k are benign skin lesions in children. The child skin lesion samples contain colour-coded identifier patches – rendering them not suitable for training in their current state. Of the remaining images, there are 3419 examples of benign skin lesions and 1082 examples of malignant skin lesions.

To make the 9k skin lesions more appropriate for training and classification, we removed the identifier patches as shown in Fig. 3. This processing step is not considered to be part of the classification framework, but rather a means to increase the amount of available training data. These identifier patches are unlikely to be present in real-world encounters. The processed child skin lesions are combined with the rest of the benign skin lesions.

The ISIC dataset [8] does not specifically provide distinctively labelled and unlabelled datasets. We partition the data into 7k unlabelled data samples, 5k labelled data samples to be used for training and the rest, 500 each for testing and validation. To expand each dataset, we performed data augmentation, by flipping the examples of skin-lesions in both the \( x \) and \( y \) (vertical and horizontal) axes and rotating the samples up to 180°. The images fed into the network are resized to have dimensions \( 64 \times 64 \times 3 \).

**Fig. 2** ssDAAE model. Image data, \( x \), is corrupted before being encoded. The encoding process consists of two sub-mappings of the corrupted image, \( \tilde{x} \), yielding an encoding of the image appearance \( \tilde{z} \) and a label prediction, \( \tilde{y} \). The decoder uses both of these to reconstruct a version of the uncorrupted image, \( \hat{x} \). The blue (striped) parts correspond to an AAE model, while the red (solid) parts are additions that make this model an ssDAAE. The label regulariser \( T' \) was first used to encourage a better classification, but its adversarial nature made the training less stable. It was used in the ablation study (4), and discarded in the reproducibility experiments (5)
3 Experiments and results

In this section, we perform an exhaustive ablation study to isolate the effectiveness of (a) a model that incorporates denoising, (b) the use of an AAE (generative model) opposed to a CNN (discriminative model) and (c) of utilising additional unlabelled data. From these experiments, we will be able to isolate exactly which components of the ssDAAE are necessary to achieve good performance. We start by explaining how the performance of our models is evaluated.

3.1 Evaluation

There is a significant label imbalance that can be observed in the ISIC dataset, meaning that the majority of the images (c. 90%) are benign. Choosing a single classification accuracy as a performance metric might be misleading given that a system that always outputs the benign class would attain, on average, a high score. Instead, we prefer to use clinically insightful and interpretable metrics, such that the percentage of malignant skin lesions correctly classified as malignant (true positive, or sensitivity) and benign skin lesions correctly classified as benign (true negative, or specificity). Furthermore, in the context of a medical application and because of the label imbalance problem, we are particularly interested in comparing model performance, in terms of specificity, at high sensitivity values, to avoid misdiagnosing a malignant skin lesion as benign.

We used similar evaluation metrics to those used in the ISIC classification challenge. Amongst the tasks of this challenge, the one most relevant to this study, skin lesion classification, had participants’ models ranked according to the specificity of their model given a particular sensitivity threshold \{0.82, 0.89, 0.95, 0.99\}.

3.2 Training and architectural details

3.2.1 Architectural details: In this subsection, we present the detailed architecture of the ssDAAE and the CNN model that was used as a baseline. The structures of these models are also summarised in Tables 2–5.

**CNN architecture:** The baseline CNN model consists of a sequence of four convolution layers, a ReLU non-linearity is applied to the output of each layer before being fed to the next one. The output of the CNN sequence is then flattened and fed to a linear layer containing 1000 neurons followed by a final linear layer, with one neuron and sigmoid non-linearity that returns values between (0, 1), where 0 is the assigned label for benign skin lesions and 1 for malignant skin lesions.

**ssDAAE architecture:** The CNN described above, without the final linear layer, forms the encoder, $E_{\theta}$ of the ssDAAE. The

1108 IET Comput. Vis., 2018, Vol. 12 Iss. 8, pp. 1105-1111 © The Institution of Engineering and Technology 2018
describe in detail how we balance the cost functions used to train both the latent variable

• cross entropy loss function was used. The sSDAAE, on the other hand, consists of several modules each with their own loss function; these loss functions need to be combined with care. The

ii. Classifier, $E^*_y$: consists of a linear layer with a single neuron followed by a sigmoid activation function. The output of this layer is the class prediction, $\hat{y}$, for a given input image.

The label output $\hat{y}$ and the latent variable $\hat{z}$ are concatenated and fed to the decoder. Two discriminators may be added to regularise the latent variable and the label (refer to Fig. 2 for a visualisation).

• Latent space discriminator, $T_{\theta^y}$: This model consists of a linear layer of 1000 neurons and a ReLU non-linearity, followed by a linear layer with a single neuron and sigmoid non-linearity.

• Binary label discriminator, $T_{\theta^y}$: This model has similar architecture to the latent space discriminator.

• Decoder, $D_{\theta^y}$: Finally, this model consists of a linear layer followed by a sequence of four transposed convolution layers, again a ReLU non-linearity is applied to the output of each transposed convolution layer before being fed to the next one. Finally, a sigmoid layer is applied to the last output of the sequence. The input to the decoder is the concatenation of the label and the encoded vector.

3.2.2 Preprocessing and input corruption: All images have been scaled to take values between −1 and 1. Furthermore, in order to allow for the partial corruption of the input, a corruption process involving additive Gaussian noise was introduced. Various $\sigma$ values between 0.0 and 1.0 were tried. Experimentally, the best results were obtained for $\sigma = 0.1$ (see Section 5.1).

3.2.3 Loss functions and class imbalance: In this section, we describe in detail how we balance the cost functions used to train our networks. For training the baseline CNN model, a single binary cross entropy loss function was used. The ssDAAE, on the other hand, consists of several modules each with their own loss function; these loss functions need to be combined with care. The loss functions include a classification loss at the output of the classifier, $E^*_{\theta^y}$, a reconstruction loss at the output of the decoder, $D_{\theta^y}$ and the regularisation loss, $T_{\theta^y}$. The latent space discriminator loss is described in (2). This cost may be modified for the label discriminator by replacing $p(z)$ with a binary distribution, and $p_g(z)$ with the output of $E^*_{\theta^y}$. We now describe the encoder loss function $L_{enc}$ defined as

\[ L_{enc} = \beta L_c + \eta L_r + \alpha (L_c + L_r) \]  

where $\alpha$, $\beta$ and $\eta$ are scalar coefficients, is the weighted combination of the following losses:

• Classification loss $L_c$: The binary cross-entropy loss between the predicted class and the ground truth label.

• Reconstruction loss $L_r$: The mean squared-error between the decoded image and the input image.

• Latent regularisation loss $L_r$: The binary-cross entropy loss between the output of the latent discriminator, $T_{\theta^y}$, and a target label. Note that a value of 1 refers to the discriminator predicting that a sample is from the chosen prior distribution.

• Label regularisation loss $L_l$: The binary-cross entropy loss between the output of the binary label discriminator, $T_{\theta^y}$, and a target label. Note that 1 refers to the discriminator predicting that a sample is from a binary distribution.

The weights $\alpha$, $\beta$ and $\eta$ are determined through experimentation (see Section 3.2.4).

Furthermore, due to the heavy class imbalance in the ISIC dataset (90% of the data is benign), it was also necessary to modify the cross-entropy loss function for the classification loss by adding a weight, $\gamma$ for label 1 and a different weight, $(1 - \gamma)$ for label 0, which leads to the following expression:

\[ L_c = \gamma y \log(\hat{y}) + (1 - \gamma)(1 - y) \log(1 - \hat{y}) \]  

3.2.4 Hyper parameter choices: For both the baseline model and our AAE model, we used $\gamma = 0.9$ for the weighted classification loss. The CNN was trained using an RMSProp optimiser with a momentum of 0 and a learning rate of $10^{-4}$. The encoder and decoder of the ssDAAE were trained with the same optimiser with same learning rate and momentum as the CNN. We found that setting coefficients $(\alpha, \beta, \eta) = (0.1, 1.0, 0.1)$ worked well. The same optimiser and learning rate were used for the regulariser, but the momentum was set to 0.2.

4 Ablation study

To appreciate the contributions of our proposed model, we performed ablation studies. We trained six different models listed

| Table 4 | Regulariser architecture |
|---------|--------------------------|
| Operation | Kernel | Strides | Padding | Filter | Non-linearity |
| (nz) latent variable input | | | | |
| linear layer | 1 × 1 | 1 × 1 | — | 1000 | ReLU |
| linear layer | 1 × 1 | 1 × 1 | — | 1 | sigmoid |
| optimiser | | | | | RMSProp ($lr = 2 \times 10^{-4}$, $m = 0.2$) |
| batch size | 64 | | | | Gaussian ($\mu = 0$, $\sigma = 1.0$), Constant (0) |

$nz$ represents the dimension of the latent variable.

| Table 5 | Label regulariser architecture for a binary classification |
|---------|--------------------------|
| Operation | Kernel | Strides | Padding | Filter | Non-linearity |
| (1,1) label input | | | | | |
| linear layer | 1 × 1 | 1 × 1 | — | 1000 | ReLU |
| linear layer | 1 × 1 | 1 × 1 | — | 1 | sigmoid |
| optimiser | | | | | RMSProp ($lr = 2 \times 10^{-4}$, $m = 0.2$) |
| batch size | 64 | | | | Gaussian ($\mu = 0$, $\sigma = 1.0$), constant (0) |
in Table 6. Each autoencoding model – consisting of an encoder and a decoder – had the same architecture and each CNN had the same architecture as the encoder. The CNN and CNN + noise models act as simple baselines that do not incorporate a generative model and are trained in a fully supervised way, not making use of any unlabelled data. The sAAE and sDAAE are fully supervised models, that do incorporate a generative model, in the form of an AAE. Finally, the ssAAE and ssDAAE are trained in a semi-supervised fashion to use both labelled and unlabelled data. All models were trained with the same amount of labelled data. The semi-supervised models are trained with the same amount of unlabelled data. To make the comparisons as fair as possible, we used the same hyper parameters [Learning rate, number of training epochs, amount of labelled and unlabelled data, loss function weightings, level of corruption, size of encoding.] for all models in the study.

The results of our ablation study are shown in Table 7. For the different sensitivity values, the ssDAAE outperformed the simple baselines, of the CNN and the CNN with added noise (CNN + noise). In this study, the ssDAAE also outperformed the ssAAE, suggesting that the corruption process is useful, but perhaps, more so in the semi-supervised model where there are more examples since the sAAE outperformed the sDAAE at the lowest sensitivity (0.82).

In addition, the CNN outperformed the CNN + noise model at all sensitivity values, further suggesting that many more training examples are needed for denoising to be effective. The fact that the CNN + noise performed less well than a CNN for all sensitivities, in contrast to the sAAE and sDAAE, which do perform well at the lower sensitivities, may be because the CNN + noise network is never exposed to the uncorrupted images, while the autoencoder models are exposed to uncorrupted images when the reconstruction loss is computed.

It is at the higher sensitivities (0.89, 0.95, 0.99) that we most clearly see that all semi-supervised variants outperformed their supervised variants. The benefits of semi-supervised models over fully supervised suggested by the results supports our motivation to design models that incorporate unlabelled data with labelled. Furthermore, the additional benefit of incorporating a denoising criterion into semi-supervised models has, as anticipated, also improved performance. Finally, our results suggest that the model that most consistently performs well is our proposed model, the ssDAAE.

5 Reproducibility experiments

CNNs are trained using a stochastic process (e.g. SGD), which means that each time a network is trained and evaluated, different results might be obtained. For many applications, it may be tempting to simply repeat the training process multiple times and use results from the best performing model. However, when we wish to compare whether a model benefits from the addition of a component, for example the addition of a denoising process, it is important to make sure that the results obtained are meaningful and not just the result of a lucky run. The results presented in the ablation study were obtained from training each model once for specific parameters settings, and not the result of choosing the best performing models from multiple runs.

To evaluate reproducibility, we re-train each of our models, the sAAE, ssAAE, sDAAE and ssDAAE eight times (32 models trained from a random initialisation) and evaluate them on the test set. This allows us to evaluate the mean and median performance associated with each model.

The mean performance of the models is shown in Fig. 4a. At the higher sensitivities the denoising semi-supervised models, on average, outperformed all other models. The benefits of using denoising and semi-supervised training, on average, are greatest at the higher sensitivities, suggesting that it is beneficial to employ both denoising and semi-supervised training.

In Fig. 4b, we show the median performance of each model. During training, we found that some model optimisations got stuck in local minima (losses remained the same for many epochs) at the start of training. Since models were trained for a fixed number of

| Table 6 | Models used for the ablation study |
|---------|----------------------------------|
| Model   | (a) Denoising | (b) Autoencoder | (c) Unlabelled |
| CNN     | ✗            | ✓              | ✓              |
| CNN + noise | ✓           | ✗              | ✓              |
| sAAE    | ✓            | ✓              | ✓              |
| sDAAE   | ✓            | ✓              | ✓              |
| ssAAE   | ✓            | ✓              | ✗              |
| ssDAAE  | ✓            | ✓              | ✓              |

The semi-supervised DAAE (ssDAAE) has three core components (a) denoising, (b) an AAE and (c) are trained in a semi-supervised fashion, training with additional unlabelled data. The sAAE and sDAAE are fully supervised models.

| Table 7 | Ablation study on the ISIC dataset |
|---------|----------------------------------|
| Model   | SP82   | SP89   | SP95   | SP99   |
| CNN     | 0.77   | 0.76   | 0.67   | 0.54   |
| CNN + noise | 0.77 | 0.67 | 0.48 | 0.21 |
| sAAE    | 0.81   | 0.72   | 0.62   | 0.43   |
| sDAAE   | 0.83   | 0.77   | 0.57   | 0.15   |
| ssAAE   | 0.80   | 0.77   | 0.72   | 0.65   |
| ssDAAE  | 0.82   | 0.79   | 0.75   | 0.68   |

The study evaluates the contributions of each component of the DAAE model. SPX values represent the specificity measured at 3% sensitivity.

Fig. 4 Reproducibility experiments with the mean and median performance. For each panel of graphs (a and b), the bars above the dashed lines indicate the specificity at four different sensitivities. The bars below the dashed lines capture AUC. These experiments have been performed without the label regulariser, $T_{\theta}$. Note that the median and mean performance of the ssDAAE is markedly better at the two highest sensitivities

(a) Mean performance over eight separately trained models. (b) Median performance over eight separately trained models
training epochs, amount of labelled and unlabelled data, loss
functions, size of encoding. To make the comparison as fair as possible, our results suggest that the optimal corruption level is $\sigma = 0.1$ for most of the sensitivity values. We see that, for all the sensitivity values, an ssDAAE trained with a noise level of $\sigma = 0.1$ outperformed an ssAE (a model trained with a noise level of $\sigma = 0$). For ssDAAE models trained with noise levels greater than $\sigma = 0.25$, inclusive, performance dropped significantly for all the sensitivity values, suggesting that too much noise may have an adverse effect on training.

6 Conclusion

Despite the clear success of deep learning techniques in specific image datasets, wide adoption of the many available approaches to training deep networks is highly dependent on the availability of sufficient quantities of [label, image] pairs.

The solution proposed in this work is a form of semi-supervised learning, in the sense that if ground truth labels are available for only a subset of the data, all the data can still be used to train a deep classification model. Our results show that the additional information that may be learned from the unlabelled data is useful for classification.

Our solution also includes a denoising procedure. While an AAE [5] is trained to simply recover its input, our model is trained to recover clean data samples from corrupted ones. This results in our model learning a robust data representation, which in turn boosts classification performance for the highest specificity.

A key feature of this work is that, unlike the case of many deep networks that have been applied to specific tasks of medical imaging, we do not use other sources of visual data for pre-training our networks. A probable consequence of this is that the resulting networks are likely to be more finely honed for the data they will be applied to. It seems evident that pre-trained networks (such as CNNs trained on ImageNet) learn features that are biased to those datasets, even after fine-tuning on specific datasets such as ISIC.

The application of AAEs is not limited specifically to the form of image data explored in this paper. They are flexible, and may potentially be applied to other datasets, where there is a large amount of image data, but a limited amount of it is labelled. The semi-supervised approach holds particular relevance for developing high specificity classification systems for existing and emerging sources of medical images. This is because it is often the case that it is very easy to collect many examples of unlabelled images, but the availability of expertise to provide ground truth labelling is limited.

7 Acknowledgments

The authors acknowledge the Engineering and Physical Sciences Research Council for funding Antonia Creswell through a Doctoral Training studentship (EP/L504786/1) and DAMAE Medical for funding Alison Pouplin. We would also like to thank Nick Pawlowski and Martin Rajchl for help with providing access to cluster computers. They also acknowledge Steff Garasto for help with figure formatting.

8 References

[1] Esteva, A., Kuprel, B., Novoa, R.A., et al.: ‘Dermatologist-level classification of skin cancer with deep neural networks’. Nature, 2017, 542, (7639), pp. 115–118
[2] Bengio, Y., Yao, L., Alain, G., et al.: ‘Generalized denoising auto-encoders as generative models’. Advances in Neural Information Processing Systems, 2013, pp. 899–907
[3] Im, D.J., Ahn, S., Memisevic, R., et al.: ‘Denoising criterion for variational auto-encoding framework’. arXiv preprint arXiv:1511.06406, 2015
[4] Kingma, D.P., Welling, M.: ‘Auto-encoding variational Bayes’. Proc. 2015 Int. Conf. Learning Representations (ICLR-2015), arXiv preprint arXiv:1312.6114, 2014
[5] Makhzani, A., Shlens, J., Jaitly, N., et al.: ‘Adversarial autoencoders’. arXiv preprint arXiv:1511.05644, 2015
[6] Vincent, P., Larochelle, H., Bengio, Y., et al.: ‘Extracting and composing robust features with denoising autoencoders’. Proc. 25th Int. Conf. Machine Learning, 2008, pp. 1096–1103
[7] Vincent, P., Larochelle, H., Lajoie, I., et al.: ‘Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion’. J. Mach. Learn. Res., 2010, 11, (Dec), pp. 3371–3408
[8] International skin imaging collaboration: Melanoma project.
[9] Codella, N.C.F., Gutman, D., Celebi, M.E., et al.: ‘Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC)’. CoRR, abs/1710.05006, 2017
[10] Menegola, A., Tavares, J., Fornaciari, M., et al.: ‘RECOD titans at ISIC challenge 2017’. CoRR, abs/1703.04819, 2017
[11] Gonzalez-Daz, I.: ‘Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions’. CoRR, abs/1705.01976, 2017
[12] Matsunaga, K., Hamada, A., Minagawa, A., et al.: ‘Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble’. CoRR, abs/1703.03108, 2017
[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: ‘Generative adversarial nets’. Advances in Neural Information Processing Systems, 2014, pp. 2672–2680
[14] Radford, A., Metz, L., Chintala, S.: ‘Unsupervised representation learning with deep convolutional generative adversarial networks’. Int. Conf. Learning Representations (ICLR, 2016), arXiv preprint arXiv:1511.06434, 2015