sensors

MDPI

*Article*

# Medical Image Classification Based on Semi-Supervised Generative Adversarial Network and Pseudo-Labelling

**Kun Liu [1], Xiaolin Ning [1] and Sidong Liu [2,*]**

[1] School of Information Engineering, Shanghai Maritime University, Shanghai 200135, China
[2] Australia Institute of Health Innovation, Macquarie University, Sydney 2113, Australia
* Correspondence: sidong.liu@mq.edu.au

**Abstract:** Deep learning has substantially improved the state-of-the-art in object detection and image classification. Deep learning usually requires large-scale labelled datasets to train the models; however, due to the restrictions in medical data sharing and accessibility and the expensive labelling cost, the application of deep learning in medical image classification has been dramatically hindered. In this study, we propose a novel method that leverages semi-supervised adversarial learning and pseudo-labelling to incorporate the unlabelled images in model learning. We validate the proposed method on two public databases, including ChestX-ray14 for lung disease classification and BreakHis for breast cancer histopathological image diagnosis. The results show that our method achieved highly effective performance with an accuracy of 93.15% while using only 30% of the labelled samples, which is comparable to the state-of-the-art accuracy for chest X-ray classification; it also outperformed the current methods in multi-class breast cancer histopathological image classification with a high accuracy of 96.87%.

**Keywords:** digital histopathology; deep learning; generative adversarial network; *k*-means clustering; medical images classification; semi-supervised learning



**Citation:** Liu, K.; Ning, X.; Liu, S. Medical Image Classification Based on Semi-Supervised Generative Adversarial Network and Pseudo-Labelling. *Sensors* **2022**, *22*, 9967. https://doi.org/10.3390/s22249967

Academic Editor: Loris Nanni

Received: 1 November 2022
Accepted: 15 December 2022
Published: 17 December 2022



## 1. Introduction

The design and use of artificial intelligence (AI), especially deep learning (DL), is driving fundamental changes in natural language processing, visual object recognition and many other domains [1]. Since AlexNet [2] won the ImageNet Challenge in 2012, DL models have dramatically improved the state-of-the-art in object detection and image classification at large scale [3]. DL also holds promises in transforming healthcare and medicine, with encouraging results recently reported in skin cancer classification [4], pneumonia detection [5], glioma prognosis [6], diabetic retinopathy detection [7], glaucoma screening [8], interstitial lung diseases classification [9], and most recently, COVID-19 assessment [10,11], etc. DL models usually require a large number of labelled samples to train; therefore, great effort has been taken to collect and label large-scale datasets, such as the ImageNet [3] and the Winograd Schema Challenge [12], and many researchers are motivated to participate in public computational challenges to take advantage of such datasets. However, it is challenging to acquire large-scale medical image datasets, as medical images usually have restricted accessibility and require clinical expertise to annotate. These limitations hamper the translation of DL models to medical image classification.

To reduce the dependence on large-scale expert-annotated medical image datasets, several unsupervised learning methods were proposed. Deep Embedding for Clustering (DEC) [13] is one of the first unsupervised methods to cluster unlabelled data, which is based on self-training. CosFace [14] used the estimated clustering uncertainty of unlabelled samples to adjust the loss function weight to reduce the overlapping-identity label noise; however, it requires balanced labelled and unlabelled samples to estimate clustering uncertainty accurately, which is a major limitation. There are a few methods based on

transfer learning [15,16] and meta-learning [17,18]. Ahn et al. [15] proposed a hierarchical unsupervised feature extractor, which has a convolutional autoencoder on top of a pre-trained convolutional neural network (CNN). Arti Pet al. [16] fine-tuned the pre-trained AlexNet [2] and GoogleNet [19] for X-ray image classification. Maicas et al. [18] designed an unsupervised pretext task for meta-learning and then trained the model for medical image classification. However, due to the lack of domain experts' input, it is difficult for these unsupervised methods to meet the high sensitivity and specificity requirements for medical applications.

The past few years have seen an emerging application of semi-supervised learning in many computer vision tasks. Semi-supervised learning methods usually require fewer expert-annotated samples (less labelling cost) and can also take advantage of a large amount of unlabelled data (more training data). In a recent survey, Van Engelen and Hoos [20] provided an overview of semi-supervised learning methods, most of which were based on a Generative Adversarial Network (GAN) [21]. GAN is a very successful unsupervised learning method for data synthesis with a wide range of applications in medical image computing [22], such as color normalisation [23], and has been used to overcome the problem of insufficiently labelled data. Odena et al. [24] developed a class-conditional GAN for image synthesis to augment training data. GAN models are usually difficult to train with known issues like mode collapse and failure to converge; therefore, variant GAN models were proposed with improved reliability. Han et al. [25] proposed a conditional GAN based on Bayesian uncertainty estimation and noise-tolerant adversarial learning, which was validated on datasets with low dimensionality demonstrating robust performance in noise resistance. Guo et al. [26] proposed a positive-unlabelled GAN (PU-GAN), which divided the generated images into positive or negative samples based on image quality to reduce the high heterogeneity in sample quality. These GAN models improved the stability and quality of the generated samples and achieved better performance than those sophisticated discriminator stabilisation methods.

Semi-supervised learning has also been used in medical image classification. In one of our recent studies, we developed a semi-supervised GAN (SSGAN) for lung X-ray classification, which only requires a small number of labelled samples [27]. This model extended the unsupervised GAN by adding an additional class of GAN-synthesised images to guide the training process. SSGAN is able to estimate the distribution of both labelled and unlabelled data so that the discriminator network, i.e., the classifier, is more robust than those trained on the labelled samples alone. We believe SSGAN can be further improved by integrating with pseudo-labelling, i.e., assigning pseudo labels to unlabelled samples based on their distances to the labelled sample cluster centres. In this study, an enhanced semi-supervised GAN with pseudo-labelling (PLAB-GAN) is proposed for medical image classification, which can not only use unlabelled data to estimate the sample distribution but also train the classifier directly.

In summary, in this study, we have made the following contributions.

- A novel GAN model based on pseudo-labelling and semi-supervised learning was proposed to optimise the use of unlabelled data in medical image classification.
- The proposed method is methodologically innovative. We used ResNet-20 to extract features from unlabelled data and further inferred their labels based on K-means clustering. We also customised the discriminator network of GAN by converting it to a multi-class classifier, which is not only able to classify if a sample is real or fake but also to predict its class. These methods effectively strengthened the effect of image features on classification, alleviated the problem of the unobvious intra-class gap and improved the accuracy of pseudo-labelling.
- We conducted extensive experiments on two benchmark datasets, including ChestX-ray14 [28] and BreakHis [29], and demonstrated that our method could improve the state-of-the-art performance of medical image classification for lung disease diagnosis using an X-ray and for breast cancer diagnosis using histopathology images.

## 2. Materials and Methods

A novel medical image classification method was proposed by integrating pseudo-labelling into semi-supervised GAN (PLAB-GAN). The overall framework of PLAB-GAN is illustrated in Figure 1. We first clustered the unlabelled samples to the cluster centres of the labelled images to estimate pseudo labels based on the CNN features extracted from the samples using a pretrained ResNet-20 network. Secondly, from each cluster, a small number of labelled data (X_lab) and a greater number of unlabelled data (X_unlab) were selected to train the discriminator/classifier, which classifies the samples into K classes (the number of classes of the real data). We further added a new class to the discriminator output for the synthetic data (X_gen) so that the synthetic images can be classified into the K+1 category (K classes for the real data and 1 pseudo class for the synthetic data). While training the PLAB-GAN, the discriminator and generator networks were alternately updated until reaching a certain number of iterations. Finally, the trained discriminator was used for medical image classification.

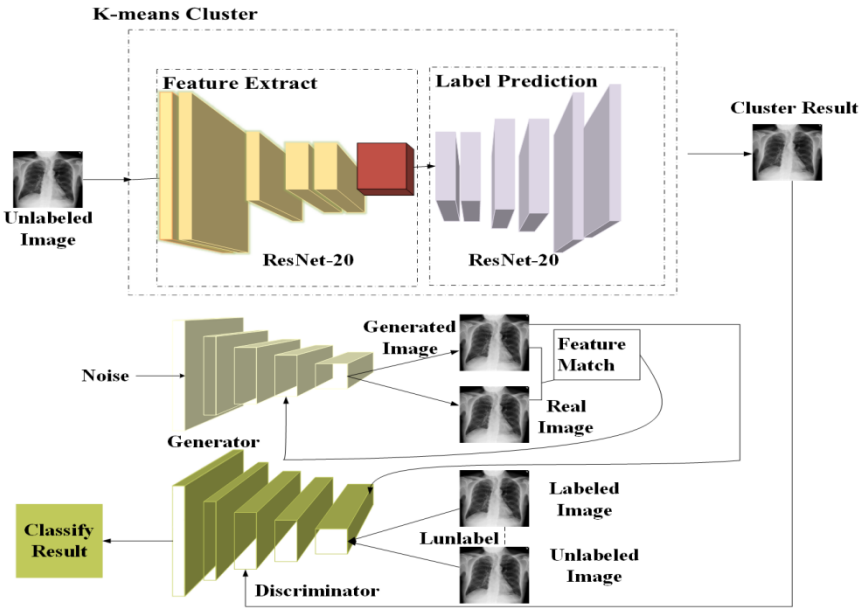


**Figure 1.** Overiew of the proposed pseudo-labelling-based semi-supervised generative adversarial network (PLAB-GAN).

### *2.1. Pseudo-Labelling Based on K-Means Clustering*

Pseudo labels, which are the estimated labels of unknown data, are generally used for processing large-scale unlabelled data. In this study, we chose K-means clustering to estimate the pseudo labels of the unlabelled images for their simplicity and robustness. Size (128 × 128 pixels) and intensity [−1,1] normalisation were first applied to the images. The preprocessed labelled images were then used to train a RestNet-20 which was pre-trained on ImageNet [3]. The network was trained through iterative learning as:

$$\text{label} = \text{argmin}\{\sum_{i=1}^{K} w^i(x^i - x_1^i)^2, \sum_{i=1}^{K} w^i(x^i - x_2^i)^2 \ldots\ldots \sum_{i=1}^{K} w^i(x^i - x_j^i)^2\} \quad (1)$$

where $x^i$ indicates the average feature vector, i.e., cluster centre, of the $i$th class ($i = \{1,\ldots,\text{K}\}$), $x_j^i$ represents the feature vector of each image, and $w^i$ express features weight. The output of activation function, $p$, which is the probability that the sample $x$ belongs to each class, was then assigned to $w^i$ to update the network weights. The last layer used Softmas as activation function, and the other layers all used ReLU.

The trained network is then used to extract features from the unlabelled samples for subsequent clustering and pseudo-labelling. For each unlabelled image, x_unlab, the

Euclidean distance between its feature f(x_unlab) and each cluster centre $x_i$ was used to estimate the pseudo label.

### 2.2. Generative Adversarial Network

A generated adversarial network (GAN) consists of a generator and a discriminator. GAN uses the idea of confrontation training which is based on game theory. A generator network G aims to produce images (x) by transforming vectors of noise z (x = G(z)) that are similar to the real images. The discriminator network D is trained to distinguish data generated from the generator distribution $p_z$ from real data. The generator network, in turn, is then trained to fool the discriminator into accepting its outputs as being real. During GAN model training, the generator G(z) and the discriminator D(x) will update their own parameters to minimise the loss. Through continuous iterative optimisation, a Nash equilibrium state—the optimal state—is finally reached by the two networks. The objective function of discriminator is defined as:

$$s^D = -E_{x-p_{data}} lb[D(x)] - E_{z-p_z} lb\{1 - D[G(z)]\} \quad (2)$$

and the objective function of the generator is defined as:

$$s^G = -E_{z-p_z} lbD[G(z)] \quad (3)$$

where $D$(*) is the discriminant probability of the discriminator, $G(z)$ is the generated image, $lb$ represents the logarithm with a base of 2, $z - p_z$ indicates the noise with random distribution, $x - p_{data}$ is the image data that follows random distribution.

### 2.3. Classification Based on GAN

The classification model we used was based on a previously proposed semi-supervised GAN (SSGAN) [27]. For a $K$-class classification problem, we added a new class to the discriminator Softmax output for the synthetic data, i.e., K classes for the real data and 1 class for synthetic data. Pseudo labels of the unlabelled images were inferred from the cluster centres of labelled samples, which were generated from K-means clustering.

The noise vectors following a normal distribution of (0, 1) were fed into the generator to generate synthetic images. The input noise vector was first converted into a one-dimensional vector by a fully connected layer (dense) and then reshaped to dimensions of $32 \times 32 \times 256$. Following the dense layer are two blocks of layers, each consisting of a 2D deconvolution layer (stride of 2 pixels), a batch norm layer, and an activation layer. For the activation function, the final output layer uses Tanh activation function, and the rest of the layers use LeakyReLU (slope on the negative half-axis was set to 0.01). Compared with the ReLU activation function, this activation function adds a linear correction unit to deal with negative input values. The size of the final generated image is $128 \times 128 \times 1$.

The generated images, a small number of labelled images and a larger number of unlabelled images were then used to train the discriminator. The discriminator network includes a conv2d layer, an activation layer (output dimensions: $64 \times 64 \times 32$), a 2nd conv2d layer, a batch norm layer, an activation layer (output dimensions: $32 \times 32 \times 64$), a 3rd conv2d layer, another batch norm layer and activation layer (output dimensions: $16 \times 16 \times 128$). The convolution layers take stride of 2 pixels with convolutional kernels of $3 \times 3$. The activation function uses LeakyReLU, with the slope on the negative half-axis set to 0.01. The last three layers of the discriminator network are a flattened layer (to convert tensor to one dimension vector), a dropout layer (to prevent overfitting) and a dense layer.

### 2.4. Loss Functions

The output of the discriminator was a K + one-dimensional logical vector, $\{l_1, l_2 \ldots l_{k+1}\}$ which was calculated by Softmax. The first K elements of the vector $(l_1, l_2 \ldots l_k)$ represent

the probabilities of being the real classes, $l_{k+1}$ represents the probability of being the synthetic class. The probability of a sample ($x$) being a specific class ($i$) can be calculated as:

$$p(y = i|x) = \frac{\exp(l_i)}{\sum_{j=1}^{k+1} \exp(l_j)} \quad (4)$$

where $\sum_{j=1}^{k+1} \exp(l_j)$ represents the sum of the probability values over the K+1 classes

The categorical cross-entropy loss was used for the labelled image classification. Binary cross-entropy was used for unlabelled images and generated images, i.e., probabilities of the sample belonging to a real class or a synthetic class. There are three types of images in the discriminator: generated images, labelled images, and unlabelled images; therefore, three types of loss functions are designed, as in Equations (5)–(7):

$$l_{label} = -E_{(x,y)-p_{data}}[lnp(y|x), y < k+1)] \quad (5)$$

$$l_{unlabel} = -E_{x-p_{data}}[\ln(1 - p(y = k+1|x)), y = k+1)] \quad (6)$$

$$l_{gen} = -E_{x-G}[lnp(y = k+1|x), y = k+1)] \quad (7)$$

where $x$ represents the image, $y$ represents the label of the image, $x$-$p_{data}$ represents the image without label and $x$-$G$ represents the generated image, ($x$,$y$)-$p_{data}$ represents the image with label, $p(|)$ indicates the predicted probability, $l_{label}$ is the cross-entropy loss of the true and the predicted class label distributions for the labelled samples, $l_{unlabel}$ is the loss for the unlabelled samples classified as a true class, and $l_{gen}$ is the loss for generated samples classified as real samples. The loss function of the discriminator ($l_d$) is the sum of $l_{label}$, $l_{unlabel}$ and $l_{gen}$, as in Equation (8), where α and β represent the weight on $l_{unlabel}$ and $l_{gen}$, respectively.

$$l_d = l_{label} + \alpha l_{unlabel} + \beta l_{gen} \quad (8)$$

The discriminator D and the generator G were trained alternatively. When training D, the weights of G were fixed, and Adam method was used to update the weights of D. Then, the weights of G were optimised by matching the features between the real and the generated images. The above steps were repeated until there was no further improvement of the model or the maximum number of iterations ($n$ = 15,000 for ChestX-ray14 and $n$ = 150 BreakHis) was reached.

## 3. Results

We tested the proposed method on two benchmark datasets, including the ChestX-ray14 dataset [28] and BreakHis [29]. All experiments were performed on a workstation with a 16GB GPU (NVIDIA, GeForce GTX1080TI). The algorithm was implemented in Python 3.6. To verify the results, we repeated the experiments 10 times and the mean accuracy values were reported. The datasets, experiments and results are described below.

### *3.1. Chest X-ray Pseudo-Labelling Results*

The ChestX-ray14 dataset contains 112,120 chest X-ray images labelled with 14 types of lung diseases. We selected seven types of common diseases and a normal control class for chest X-ray classification, as shown in Figure 2. In the *K*-means clustering, we used 16,089 labelled samples (~2100 from each class, with a train:test ratio of 7:3) to train the ResNet-20 network, which was later used to extract features from unlabelled X-ray images. Then *K*-means clustering was used to infer the pseudo labels of 8583 unlabelled samples. To train the semi-supervised GAN for classification, we used a small number of labelled images (n = 50 to 400 per class) and a large number of unlabelled images (n = 8583). The X-ray dataset was divided into training:test:validation split was set to 7:2:1. The GAN learning rate is 0.0001 and the batch size is set to 16. The semi-supervised experiment uses Adam to optimise the loss function with a momentum of 0.5. We used accuracy as the

metric to validate the effectiveness of the method, as in Equation (9), where TP is True positive, FN is False Negative, TN is True Negative, and FP is False Positive.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

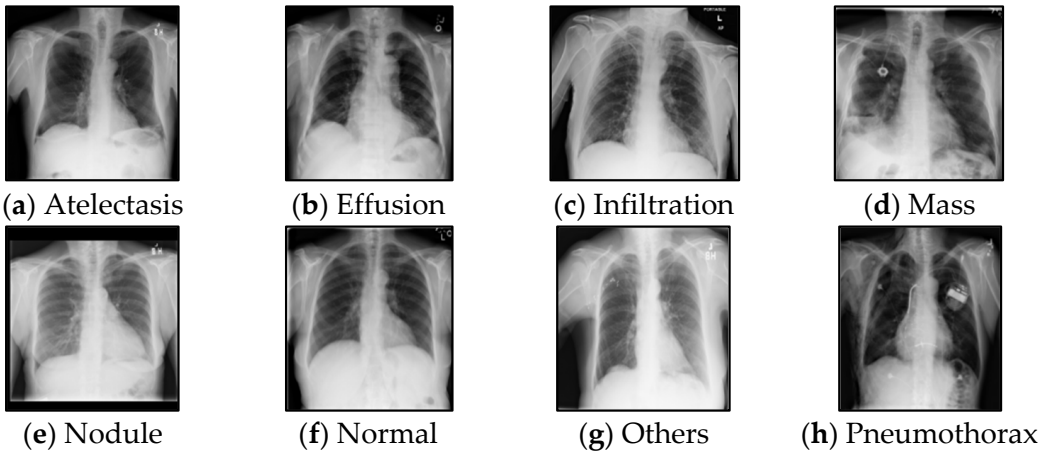


**Figure 2.** The chest X-ray sample images.

Figure 3 shows the visual representations of K-means clustering results. Figure 4 shows the experimental results using 400 labelled images in each class: changes in the accuracy (left) and the loss (right) of the model during the training process. It can be seen that the classification accuracy reached $0.860 \pm 0.026$ after 8000 epochs and further improved to $0.930 \pm 0.032$ after 15,000 epochs. The discriminator loss decreased continuously, whereas the generator loss quickly decreased in the early stage and then increased slightly later. The discriminator loss was lower than the generator loss, indicating the discriminator could distinguish the generated images very well.

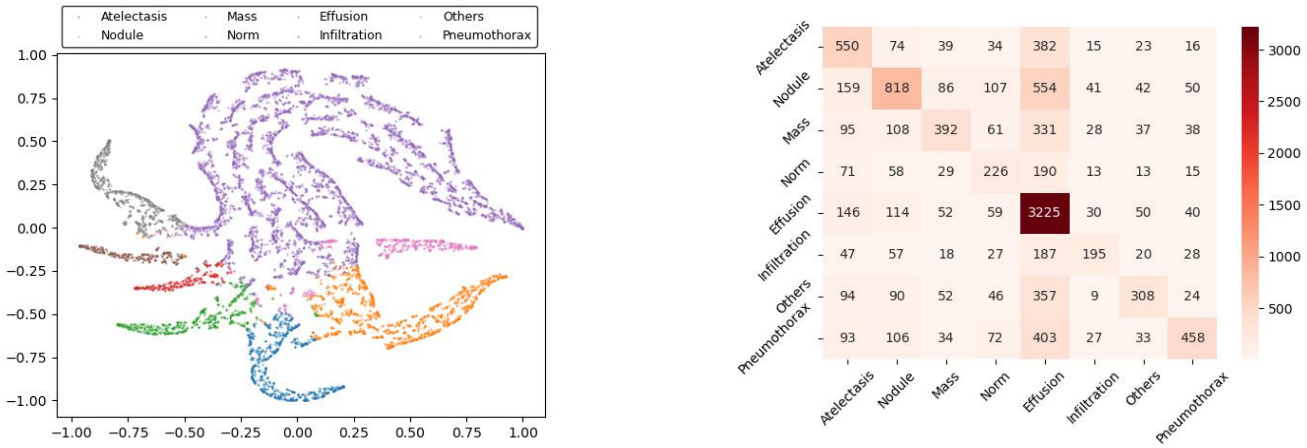


**Figure 3.** The visual representation of K-means clustering results: feature distribution after clustering (**left**); distances between the predicted label and the real label (**right**).

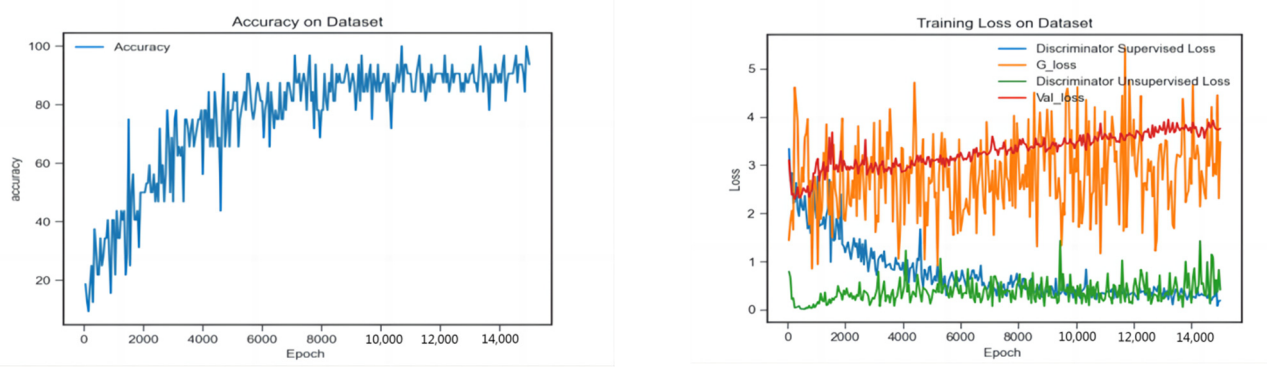


**Figure 4.** The model's performance, accuracy (**left**) and training loss (**right**) on the X-ray dataset.

### *3.2. Chest X-ray Classification Results*

To verify the effectiveness of the proposed method, we compared it with convolutional neural network (CNN), PCA-based semi-supervised method (PCA + SVM) and GAN-based semi-supervised method (SSGAN). To investigate the impact of the amount of the labelled samples for training, the experiments were repeated five times with different settings in terms of the number of images per class for training, i.e., 50, 100, 200, 300 and 400, respectively. Table 1 shows the classification accuracy with different numbers of labelled images using different networks. Compared to CNN, PCA + SVM and SSGAN, the proposed method achieved substantially better performance. Increasing the number of labelled training images improves all the models' performance. The largest performance gain was seen when using 400 labelled samples per class for training. The proposed method outperformed CNN, PCA + SVM and SSGAN by 18%, 20% and 16%, respectively.

**Table 1.** Classification accuracy of samples with different numbers of labelled data.

| No. of Labelled Samples in Each Class | Accuracy (%) | | | |
|---|---|---|---|---|
| | PCA + SVM | CNN | SSGAN | The Proposed |
| 50 | 58.94 ± 6.3 | 55.62 ± 6.2 | 62.47 ± 4.7 | **70.60 ± 5.0** |
| 100 | 63.20 ± 4.1 | 61.64 ± 5.5 | 68.71 ± 4.3 | **73.84 ± 3.2** |
| 200 | 67.76 ± 4.6 | 68.89 ± 2.0 | 72.40 ± 4.0 | **78.69 ± 4.8** |
| 300 | 68.60 ± 5.6 | 72.85 ± 5.0 | 74.25 ± 3.5 | **84.92 ± 2.6** |
| 400 | 70.10 ± 3.9 | 74.54 ± 3.4 | 77.83 ± 3.8 | **93.15 ± 3.2** |

* Bold type represents the best result.

Table 2 shows the classification accuracy (with 400 labelled training images per class) of CNN, SSGAN and the proposed method in individual classes. The proposed method outperformed SSGAN and CNN in five out of six classes, except for the Mass class.

**Table 2.** Classification accuracy of different models for individual classes.

| Class | Accuracy (%) | | |
|---|---|---|---|
| | CNN | SSGAN | The Proposed |
| Atelectasis | 79.4 | 81.97 | **94.0** |
| Nodule | 75.2 | 77.38 | **96.0** |
| Mass | **88.4** | 85.32 | 84.0 |
| Effusion | 88.2 | 82.75 | **91.0** |
| Infiltration | 70.5 | 71.62 | **86.0** |
| Pneumothorax | 87.8 | 83.41 | **93.0** |

* Bold type represents the best result.

To investigate the impact of loss weight parameters $\alpha$ and $\beta$ on the model's performance, we tested different parameter settings (values ranging from 0.1 to 0.9). Figure 5

shows the corresponding classification accuracy in different settings. It shows that when $\alpha$ and $\beta$ both were equal to 0.5, the model achieved the highest classification accuracy.

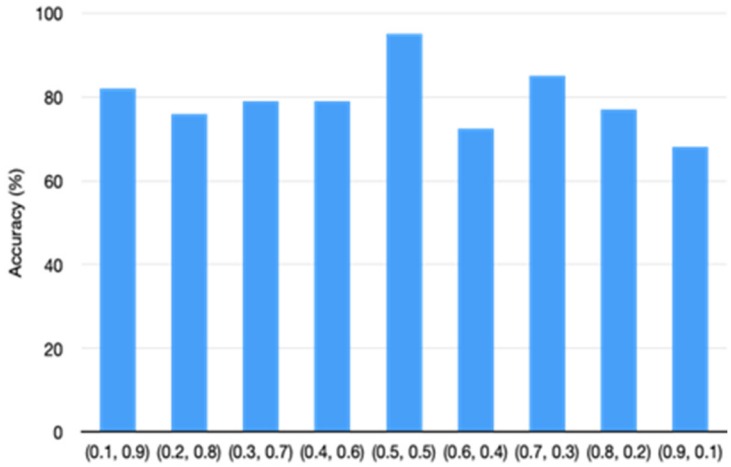


**Figure 5.** Classification accuracy with different $(\alpha, \beta)$ values.

### *3.3. BreaKHis Pseudo-Labelling Results*

We tested the proposed method on a second benchmark dataset–BreaKHis [29], which contains 7909 breast tissue microscopic images, including 2480 benign and 5429 malignant samples across eight sub-types (benign subtypes: adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma; malignant subtypes: ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma), as shown in Figure 6. The images were acquired using an Olympus BX-50 system microscope and a relay lens with a magnification of 3.3× attached to a Samsung digital color camera SCC-131AN. The images were in the three-channel RGB true color space (8 bits per channel) with different magnifications (40×, 100×, 200×, 400×). In the experiment, we expanded the dataset to 14,523 through rotation and translation operations.

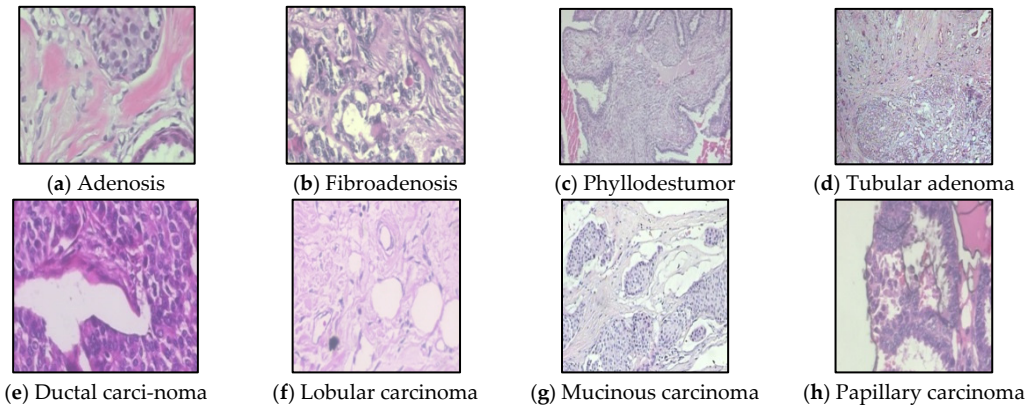


**Figure 6.** Breast cancer histopathology images from the BreakHis dataset.

To run K-means clustering, we selected 3191 labelled breast cancer images (~400 samples per class) to train the ResNet-20 network with a train:test ratio of 7:3. Image features were then extracted by the trained ResNet-20 for subsequent K-means clustering and pseudo-labelling of a total of 10,162 unlabelled breast cancer images. The same network structure, parameters and evaluation metrics as for the X-ray classification experiment were used in this experiment.

Figure 7 shows the classification performance on the BreakHis dataset. It shows that the classification accuracy was the highest (0.9687) after 140 epochs. From Figure 7 (right),

it can be seen that the discriminator loss gradually decreased during the training; the discriminator loss was lower than the loss of the generator, indicating that the discriminator performed well in recognition of the labelled, unlabelled and synthetic samples.

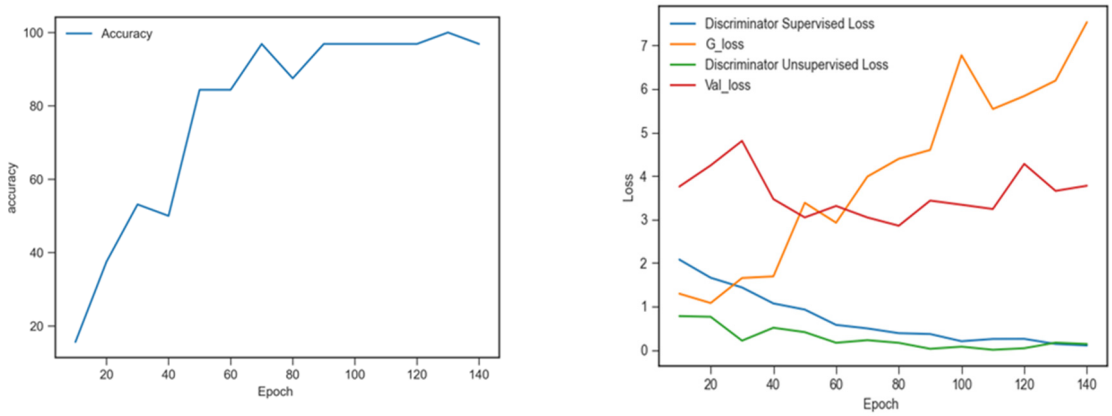


**Figure 7.** Classification performance on BreakHis dataset: accuracy (**left**), training loss (**right**).

### *3.4. BreaKHis Classification Results*

For the breast cancer image classification experiment, three different settings in terms of the number of labelled training samples were tested, i.e., 10, 20 and 30, respectively. Table 3 shows the accuracy of the model compared to CNN and SSGAN. We also compared our results with the recently published results on the same dataset, including ResNet50 [30,31] and three other types of CNN/DNN models [32–35]. The comparison with these methods is shown in Table 4. Table 5 further shows the classification accuracy in each individual subtype. As can be seen from Tables 3–5 the proposed algorithm achieved very good overall classification performance (96.87%) and also consistently high performance (95.60–97.31%) across different subtypes while using only a small number of labelled samples. It also outperformed the state-of-the-art methods.

**Table 3.** Classification accuracy of samples with different numbers of labelled data.

| Labelled Data | Accuracy (%) | | |
|---|---|---|---|
| | **CNN** | **SSGAN** | **The Proposed** |
| 10 | 68.87 | 75.37 | $95.10 \pm 0.20$ |
| 20 | 72.35 | 81.36 | $96.00 \pm 0.70$ |
| 30 | 73.68 | 84.63 | $96.87 \pm 0.50$ |

**Table 4.** Classification accuracy of different models.

| Model | Accuracy (%) |
|---|---|
| Y. Yari et al. [30] | 93.35 |
| M. Nawaz et al. [31] | 95.00 |
| P. Nguyen et al. [32] | 73.68 |
| S. Pratiher et al. [33] | 95.46 |
| D. Bardou et al. [34] | 88.23 |
| Z. Han et al. [35] | 93.80 |
| The Proposed | 96.87 |

**Table 5.** The classification accuracy of each type of breast cancer disease data.

| Major Class | Subclass | Accuracy (%) |
|---|---|---|
| benign | adenosis | 96.12 |
| | fibroadenoma | 96.88 |
| | phyllodes tumor | 96.74 |
| | tubular adenoma | 95.60 |
| malignant | ductal carcinoma | 97.31 |
| | lobular carcinoma | 96.80 |
| | mucinous carcinoma | 95.78 |
| | papillary carcinoma | 96.87 |

## 4. Discussion

A bottleneck exists in supervised learning for medical image classification. It is difficult to obtain a large number of labelled medical images for training due to the restriction in accessing, sharing and labelling patients' data. Developing robust and effective DL models with limited labelled data remains a major challenge in computer vision tasks, including medical image classification. To address this challenge, we proposed a novel method that used pseudo-labelling and semi-supervised GAN to classify medical images. This method effectively reduced the dependence of DL models on large-scale labelled data. The technical innovation of our method firstly includes our training of a ResNet-20 model to extract image features from the medical images, which could robustly assign pseudo labels to unlabelled images; secondly, our method enforcing the similarity between similar image features and minimising the intra-class distances, which strengthened pseudo-labelling performance and also improved the characterisation of image features effectively.

While DL will be increasingly used in medical image classification, how to use a large amount of information in unlabelled medical images is an emerging research area. Our proposed method provides a feasible solution for medical image classification, which uses a small amount of labelled data but can achieve equivalent or better performance compared to supervised learning methods. Our study demonstrates that pseudo-labelling and semi-supervised GAN might be a good option for the future development of intelligent medical image classification systems.

One limitation of the proposed method is that the quality of the ground truth labels has a huge impact on the clustering/pseudo-labelling, as well as the subsequent classification performance. If the ground truth labels are incorrect, the pseudo labels based on them will become less reliable; thus, the errors will propagate to the feature extractor and classifier, leading to misclassification and lower classification accuracy. Potential solutions to this problem include a mechanism to detect out-of-distribution samples [36], such as anomalies and adversarial samples, from the training set and novel ambiguity quantification functions [37] to regulate the weights of unreliable training samples. This will be investigated in our future studies.

## 5. Conclusions

To improve the classification accuracy of medical images and reduce the use of labelled images, we proposed a novel method based on *K*-means clustering/pseudo-labelling and semi-supervised GAN. Comprehensive experiments were carried out on two benchmark datasets, including ChestX-ray14 and BreakHis. The results demonstrate that our algorithm outperformed the state-of-the-art methods and worked effectively in medical image classification with a small number of labelled samples. It achieved 93.15% accuracy in X-ray classification with 400 labelled images per class and 96.87% accuracy in breast histopathology image classification with only 30 labelled images per class. The method has a high potential to assist in tasks where the unlabelled data is rich, but the labelling cost is high. In our future studies, we will further investigate novel strategies to enhance the model's performance and robustness.

**Author Contributions:** Methodology, K.L.; software, X.N.; writing—original draft preparation, X.N. and K.L.; writing—review and editing, S.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funding by the Natural Science Foundation of China (Grant No. 61803257) and the Aeronautical Science Foundation of China under Grant 201955015001.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Two public benchmark datasets were used in this study, including the ChestX-ray14 dataset [28] and BreakHis [29].

**Acknowledgments:** This work is sponsored by the Natural Science Foundation of China (Grant No. 61803257) and the Aeronautical Science Foundation of China under Grant 201955015001.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
2. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
3. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
4. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
5. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
6. Liu, S.; Shah, Z.; Sav, A.; Russo, C.; Berkovsky, S.; Qian, Y.; Coiera, E.; Dileva, A. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. *Sci. Rep.* **2020**, *10*, 7733. [CrossRef]
7. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef]
8. Liu, S.; Graham, S.; Schulz, A.; Kalloniatis, M.; Zangerl, B.; Cai, W.; Gao, Y.; Chua, B.; Arvind, H.; Grigg, J.; et al. A Deep Learning-based Algorithm Identifies Glaucomatous Discs using Monoscopic Fundus Photographs. *Ophthalmol. Glaucoma* **2018**, *1*, 15–22. [CrossRef]
9. Anthimopoulos, M.; Christodoulidis, S.; Ebner, L.; Christe, A.; Mougiakakou, S. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Trans. Med. Imaging* **2016**, *35*, 1207–1216. [CrossRef]
10. Quiroz, J.C.; Feng, Y.Z.; Cheng, Z.Y.; Rezazadegan, D.; Chen, P.-K.; Lin, Q.-T.; Qian, L.; Liu, X.-F.; Berkovsky, S.; Coiera, E.; et al. Development and Validation of A Machine Learning Approach for Automated Severity Assessment of COVID-19 based on Clinical and Imaging Data. *JMIR Med. Inform.* **2021**, *9*, e24572. [CrossRef]
11. Feng, Y.Z.; Liu, S.; Cheng, Z.Y.; Quiroz, J.C.; Rezazadegan, D.; Chen, P.-K.; Lin, Q.-T.; Qian, L.; Liu, X.-F.; Berkovsky, S.; et al. Severity Assessment and Progression Prediction of COVID-19 Patients based on the LesionEncoder Framework and Chest CT. *Information* **2021**, *12*, 471. [CrossRef]
12. Levesque, H.J.; Davis, E.; Morgenstern, L. The Winograd Schema Challenge. In Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning, Rome, Italy, 10–14 June 2012; pp. 552–561.
13. Xie, J.; Girshick, R.B.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015.
14. RoyChowdhury, A.; Yu, X.; Sohn, K.; Learned-Miller, E.; Chandraker, M. Improving Face Recognition by Clustering Unlabelled Faces in the Wild. In Proceedings of the ECCV, Glasgow, US, 23-28 August 2020.
15. Ahn, E.; Kumar, A.; Feng, D.; Fulham, M.; Kim, J. Unsupervised Deep Transfer Feature Learning for Medical Image Classification. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019.
16. Arti, P.; Agrawal, A.; Adishesh, A.; Lahari, V.M.; Niranjana, K.B. Convolutional Neural Network Models for Content Based X-ray Image Classification. In Proceedings of the 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Bangalore, India, 15–16 November 2019.
17. Hsu, K.; Levine, S.; Finn, C. Unsupervised Learning via Meta-Learning. In Proceedings of the ICLR, New Orleans, LA, USA, 6–9 May 2019.
18. Maicas, G.; Nguyen, C.; Motlagh, F.; Nascimento, J.C.; Carneiro, G. Unsupervised Task Design to Meta-Train Medical Image Classifiers. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020.

19. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
20. van Engelen, J.E.; Hoos, H.H. A Survey on Semi-Supervised Learning. *Mach. Learn.* **2019**, *109*, 373–440. [CrossRef]
21. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the NIPS, Montreal, Canada, 8-13 December 2014; pp. 2672–2680.
22. Jose, L.; Liu, S.; Russo, C.; Nadort, A.; di Ieva, A. Generative Adversarial Networks in Digital Pathology and Histopathological Image Processing: A Review. *J. Pathol. Inform.* **2021**, *12*, 43. [CrossRef] [PubMed]
23. Cong, C.; Liu, S.; di Ieva, A.; Pagnucco, M.; Berkovsky, S.; Song, Y. Colour Adaptive Generative Networks for Stain Normalisation of Histopathology Images. *Med. Image Anal.* **2022**, *82*, 102580. [CrossRef] [PubMed]
24. Odena, A.; Olah, C.; Shlens, J. Conditional Image Synthesis with Auxiliary Classifier GANs. *arXiv* **2017**, arXiv:1610.09585v4.
25. Han, L.; Gao, R.; Kim, M.; Tao, X.; Liu, B.; Metaxas, D. Robust Conditional GAN from Uncertainty-Aware Pairwise Comparisons. *arXiv* **2019**, arXiv:1911.09298v1. [CrossRef]
26. Guo, T.; Xu, C.; Huang, J.; Wang, Y.; Shi, B.; Xu, C.; Tao, D. On Positive-Unlabelled Classification in GAN. *arXiv* **2020**, arXiv:2002.01136v1.
27. Liu, K.; Wang, D.; Rong, M. X-ray image classification algorithm based on semi-supervised generative adversarial network. *Acta Opt. Sin.* **2019**, *39*.
28. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3462–3471.
29. Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 1455–1462. [CrossRef]
30. Yari, Y.; Nguyen, H.; Nguyen, T.V. Accuracy Improvement in Binary and Multi-Class Classification of Breast Histopathology Images. In Proceedings of the 2020 IEEE Eighth International Conference on Communications and Electronics (ICCE), Phu Quoc Island, Vietnam, 13–15 January 2021.
31. Nawaz, M.A.; Sewissy, A.A.; Soliman, T.H.A. Automated Classification of Breast Cancer Histology Images using Deep Learning based Convolutional Neural Networks. *Int. J. Comput. Sci. Netw. Secur.* **2018**, *4*, 152–160.
32. Nguyen, P.T.; Nguyen, T.T.; Nguyen, N.C.; Le, T.T. Multiclass Breast Cancer Classification Using Convolutional Neural Network. In Proceedings of the 2019 International Symposium on Electrical and Electronics Engineering (ISEE), Ho Chi Minh City, Vietnam, 10–12 October 2019.
33. Pratiher, S.; Chattoraj, S. Diving Deep onto Discriminative Ensemble of Histological Hashing & Class-Specific Manifold Learning for Multi-class Breast Carcinoma Taxonomy. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019.
34. Bardou, D.; Zhang, K.; Ahmad, S.M. Classification of Breast Cancer based on Histology Images using Convolutional Neural networks. *IEEE Access* **2018**, *6*, 24680–24693. [CrossRef]
35. Han, Z.; Wei, B.; Zheng, Y.; Yin, Y.; Li, K.; Li, S. Breast Cancer Multi-Classification from Histopathological Images with Structured Deep Learning Model. *Sci. Rep.* **2017**, *7*, 4172. [CrossRef] [PubMed]
36. Raghuram, J.; Chandrasekaran, V.; Jha, S.; Banerjee, S. A General Framework for Detecting Anomalous Inputs to DNN Classifiers. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021.
37. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R.; et al. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Inf. Fusion* **2021**, *76*, 243–297. [CrossRef]