A Comparative Study of Medical Image Classification on a Small Dataset

1st Jianbing Li
Faculty of Business and Information Technology
Whitireia Community Polytechnic
Auckland, New Zealand
ljb_wh@hotmail.com

***Abstract*—Medical image classification plays a vital role in clinical treatment and teaching tasks. However, the traditional method has reached its ceiling on performance. Moreover, by using them, much time and effort needs to be spent on extracting and selecting classification features. The deep neural network is an emerging machine learning method that has proven its potential for varying classification tasks. Especially, the convolutional neural network dominates with the best results on varying image classification tasks. However, medical image datasets are hard to collect because it needs a lot of professional expertise to label them. Therefore, this report researches how to apply the convolutional neural network-based algorithm on a small chest X-Ray dataset to classify pneumonia. Totally, three techniques are evaluated through experiments. These are Linear Support Vector Machine classifier with local rotation and orientation free features, transfer learning on two Convolutional neural network models: VGG16 and InceptionV3, and a capsule network training from scratch. Data augmentation as a data preprocessing method are applied to all three methods. The results of the experiments show that data augmentation generally is an effective way for all three algorithms to improve performance; Transfer learning is a more effective classification method on a small dataset compared to SVM with ORB and Capsule Network; In transfer learning, retraining specific features on a new target dataset is essential to improve performance and the second important factor is a proper network complexity that matches the scale of the dataset. Moreover, dropout layers with proper dropout rate, a global average pooling layer, early stopping and batch normal layers are also essential to prevent overfitting and therefore to improve the final performance.**

***Index Terms*—CNN, Transfer Learning, Capsule Network, ORB, SVM, Image Classification**

I. Introduction

Effectively classifying medical images play an important role aiding clinical care and treatment. For example, Analysis X-Ray is the best approach to diagnose pneumonia [1] which causes about 50,000 people to die per year in the US [2], but classifying pneumonia from chest X-Rays needs professional radiologists which is a rare and expensive resource for some regions.

Use of the traditional machine learning methods, such as Support Vector Methods (SVMs), in medical image classification, began long ago. However, these methods have the following disadvantages: the performance is far from the practical standard, and the developing of them is quite slow in recent years; the feature extracting and selection are time-consuming and vary according to different objects [3]. The deep neural networks (DNN), especially the Convolutional neural networks (CNN) are widely used in varying image classification tasks and have achieved significant performance since 2012 [4]. Some research on medical image classification by CNN has achieved performances rivalling human experts. For example, CheXNet, a CNN with 121 layers trained on a dataset with more than 100,000 frontal-view chest x-rays (ChestX-ray 14), achieved a better performance than the average performance of four radiologists. Moreover, Kermany et al. [3] propose a transfer learning system to classify 108,309 Optical coherence tomography (OCT) images, and the weighted average error is equal to the average performance of 6 human experts.

The medical images are hard to collect, as the collecting and labelling of medical data are confronted with both data privacy issues and the need for time-consuming expert annotations. In the two general resolving directions, one is to collect more data, such crowdsourcing [5] or digging into the existing clinical reports [6]. Another way is studying how to improve the performance on a small dataset, which is very important because the knowledge achieved from the research can migrate to the research on big datasets. In addition to this the most significant published Chest X-Ray image dataset (ChestX-ray 14) is still far smaller than the biggest general image dataset-ImageNet which has reached 14,197,122 instances at 2010 [7] [8]

The research questions of this report are:

- How can data mining, including traditional ways and CNN-based classification be used to classify small medical images dataset?
- How do the performance of different data mining methods compare with each other?

CNN-based methods have several methods to improve the performance of image classification on small datasets: One method is data augmentation [9] [10] [11] [12]. Wang and Perez [13] researched the effectiveness of Data Augmentation in image classification. The authors found the traditional transform-based Data Augmentation have better performance than GAN and other neural network-based methods. Another method is transfer learning [14] [3] [12] [15]. Kermany et al. [3] achieved 92% accuracy on a small pneumonia X-rays image dataset by transfer learning. The third method is the Capsule Network. Sabour et al. [16] invented a new neural

network structure-capsule network, which achieves state-of-the-art performance on the Modified National Institute of Standards and Technology (MNIST) database [17] and also the best performance on other small datasets. Afshar et al. [18] have utilised Capsule network to detect brain tumours and got 86.56% accuracy.

However, there are some gaps needing to be noticed. A limitation of Kermany's research is they use InceptionV3 model and stop retrain the convolutional layer of InceptionV3 because of the overfitting. Therefore, other models and the effects of retraining the convolutional layer will be evaluated in this research. Moreover, Afshar et al. [18] did not compare the performance of capsule network with other methods. Therefore, the contributions of this report include:

- A performance comparison of three different classification methods: SVM classifier with Oriented fast and Rotated Binary robust independent elementary features (ORB), transfer learning of VGG16 and InceptionV3 and training capsule network from scratch.
- An analysis of the effects of data augmentation, network complexity, fine-tuned convolutional layer and other preventing overfitting mechanics on the classification of small chest x-ray dataset by transfer learning of CNN.

This report conducts four groups of experiments. Except the SVM with ORB is running on a standard laptop, Convolutional neural network related experiments are all run on a virtual machine with a Nvidia Tesla K80 Graphic card in Google Cloud [19].

The rest of this report is organized as follows: Section two reviews the related literature on medical image classification. Section three describes the design of experiments. Section four presents the result of the experiments. Section six discusses the results. Finally, the conclusion is drawn, and the future work described, followed by references.

II. Literature Review

A. Introduction

Medical image classification is a sub-subject of image classification. Many techniques in image classification can also be used on it. Such as many image enhanced methods to enhance the discriminable features for classification [20]. However, as the CNN is an end to end solution for image classification, it will learn the feature by itself. Therefore, the literature about how to select and enhance features in the medical image will not be reviewed. The review will mainly focus on the application of the traditional method, CNN-based transfer learning and capsule network on medical image related paper to investigate what factors in those models are important to the final result and the gaps they haven't included in their work.

B. ORB and SVM Application on Medical Image Classification

Paredes et al. [21] use small patches of medical images as local features and k-nearest neighbour(k-NN) to classify the categorization of the whole medical image, finally achieving start-of-art accuracy. Parveen & Sathik [22] researched to detect Pneumonia from X-Rays. The authors extracted features by Discrete Wavelet Transform (DWT), Wavelet Frame Transform (WFT) and Wavelet Packet Transform (WPT) and used Fuzzy C-means to detect Pneumonia. Caicedo et al [23] use Scale-Invariant Feature Transform (SIFT) as a local feature descriptor and use Support Vector Machines (SVM) classifiers to classify medical images and get state-of-art precision at 67%. However, SIFT is a patent algorithm. Thus, Rublee et al. [24] propose a free, faster local feature descriptor-Oriented fast and Rotated Binary robust independent elementary features(ORB), which has the same performance as SIFT and even better performance than SIFT under some condition. SVM is also a high-performance classification algorithm, widely used in different medical image classification tasks by other researchers and achieve an excellent performance [25] [26]. Therefore, this report uses ORB and SVM as the representation of the traditional methods.

C. CNN on Medical Image Classification

With the different CNN-based deep neural networks developed and achieved significant result on ImageNet Challenger, which is the most significant image classification and segmentation challenge in the image analysing field [27]. The CNN-based deep neural system is widely used in the medical classification task.

CNN is an excellent feature extractor, therefore utilizing it to classify medical images can avoid complicated and expensive feature engineering. Qing et al. [28] presented a customized CNN with shallow ConvLayer to classify image patches of lung disease. The authors also found the system can be generalized to other medical image datasets. Moreover, other research also found CNN based system can be trained from big Chest X-Ray (CXR) film dataset and got start-of-art high accuracy and sensitivity results on their dataset, like Stanford Normal Radiology Diagnostic Dataset containing more than 400,000 CXR and a new CXR database(ChestX-ray8), which consist of 108,948 frontal-view CXR [29]

Moreover, using limited data makes it hard to train an effective model, and therefore the transfer learning of CNN be wildly used in medical image classification tasks. Kermany et al. [3] use InceptionV3 with ImageNet trained weight and transfer learning on a medical image dataset containing 108,312 optical coherence tomography(OCT) images. They got an average accuracy of 96.6%, with a sensitivity of 97.8% and a specificity of 97.4%. The authors also compared the results with six human experts. Most of the experts got high sensitivity, but low specificity while the CNN-based system got high values on both sensitivity and specificity. Moreover, on the average weight error measure, the CNN-based system exceed two human experts. The authors also verified their system on a small pneumonia dataset including about 5 thousand images, and achieved an average accuracy of 92.8%, with a sensitivity of 93.2% and a specificity of 90.1%. This system finally may help in accelerating diagnosis and referral of patients and therefore introduce early treatment, resulting in

increased cure rate. Moreover, Vianna [30] also studied how to utilize transfer learning to build an X-Ray image classification system that is the critical component of a Computer-Aided-Diagnosis system. The authors found a fine-tuned transfer learning system with data augmentation effectively alleviate overfitting problem and yield a better result than two other models: training from scratch and a transfer learning model with only a retrained last classification layer.

D. Capsule Neural Network on Medical Image Classification

As mentioned in the previous section, the CapsNet was invented in 2017 [16]. Therefore, the research about it is not as fruitful as CNN. However, there still are some research on applying them to the different datasets and varying fields due to its excellent feature – Equivariance which means the spatial relationship of objects in an image is kept and at the same time the result does not impact by the object's orientation and size.

Afshar et al. [18] applied CapsNet to classifying brain tumours on Magnetic Resonance Imaging (MRI) images, and got 86.56% prediction accuracy with a modified CapsNet that reduces the feature maps from the original 256 to 64. Moreover, Tomas and Robertas [31] presented a CapsNet based solution to classify four types of breast tissue biopsies from breast cancer histology images. They achieved 87% accuracy with the same high sensitivity. Jiménez-Sánchez et al. [5] evaluated the CapsNet on medical image challenges. The authors selected a CNN with three layers of ConvLayer as the baseline and compared CapsNet's performance with LeNet and the baseline on four datasets, MNIST, Fashion-MNIST, mitosis detection (TUPAC16) and diabetic retinopathy detection (DIARETDB1), with three conditions: the partial subset of the dataset, the imbalanced subset of the dataset and data augmentation. The final result shows CapsNet performed better than the other two networks in a small, imbalanced dataset. Beşer, Kizrak, Bolat, & Yildirim [32] implemented a sign language recognizing system by CapsNet and achieved 94.2% validation accuracy.

Moreover, some research studied the internal mechanics by varying network structures under different conditions. Xi, Bing, & Jin [33] studied the impact of different network structures on a complex dataset CIFAR10. The authors choose following options:

1) Increase the number of Primary Capsule Layers.
2) Increase the capsule number in Primary Capsule Layer.
3) Assemble multiple models and average the result.
4) Adjust the scaling factor of reconstruction loss.
5) Add more ConvLayer
6) Evaluate other activation function

Finally, the authors found more ConvLayers and more models assembled have more effect on improving the final accuracy, and they achieved the highest result with a 7-model assembled CapsNet with a more ConvLayer than the original version of Sabour's. Moreover, The CapsNet of Tomas and Robertas used to classify breast cancer increased the ConvLayer to five layers. On the other hand, Afshar et al. [18] also evaluated the different options of CapsNet. They fine-tuned the input size, number of feature maps, number of ConvLayers, capsule number in primary CapsLayer, dimension number in Primary Capsule and the neuron number in reconstruction layers. The authors got the best results with a CapsNet having 64x64 input image (original is 28x28) and fewer feature map, which reduces to 64 from the original 256. Also, authors found increasing the routing iteration number beyond three will not increase the performance on the four datasets: MNIST, Fashion-MNIST, The Street View House Numbers(SVHN) dataset and Canadian Institute for Advanced Research 10 (CIFAR10) dataset.

E. Conclusion

From the previous reviews, it can be seen that the traditional method (SVM with ORB feature), CNN based transfer learning and Capsule network can all use on the medical image dataset. Just looking at the value of accuracy on different datasets, CNN based transfer learning looks have better performance than the other two methods. However, they have not been compared on the same dataset. Therefore, this report will compare their performance on the same dataset-the pneumonia dataset.

Moreover, there are so many different options when fine-tuning the parameter of those methods. The traditional method has so many features and classifying algorithms which can be evaluated. They cannot be iterated in this report due to the limited time. As the baseline, the traditional method choose ORB as the feature and linear SVM as the classifier. As the data augmentation is a data preprocessing method that can apply to all three methods, it also will be evaluated on the traditional method. For CNN-based transfer learning, the layers of retrained ConvLayer, the complexity of classification layers, the dropout rate has significant effects on the final result. Therefore, they will be evaluated by this research. Based on the same research, the critical fact in capsule network: the number of feature map, the number of the capsules and the channels of capsule will also be evaluated in this report.

III. Experiments Design

A. Data Source and Initial Analysis

The Dataset comes from the work of Kermnay et al. [34]. It contains two kinds of chest X-Ray Images: NORMAL and PNEUMONIA, which are stored in two folders. In the PNEUMONIA folder, two types of specific PNEUMONIA can be recognized by the file name: BACTERIA and VIRUS. Table I describes the Composition of the dataset. The training dataset contains 5232 x-ray images while the testing dataset contains 624 images. In the training dataset, the image in the NORMAL class only occupies one-fourth of all data. In the testing dataset, the PNEUMONIA consists of 62.5% of all data, which means the accuracy of the testing data should higher 62.5%.

Figure 1 shows examples of chest X-Rays from the dataset. The normal chest X-ray (left panel) depicts clear lungs without any areas of abnormal opacification in the image. Bacterial

TABLE I
THE COMPOSITION OF CHEST X-RAY DATASET

	Training dataset	Testing dataset
NORMAL	1349(25.7%)	234(37.5%)
BACTERIA	2538(48.5%)	242(38.7%)
VIRUS	1345(25.7%)	148(23.7%)
Total	5232(100%)	624(100%)

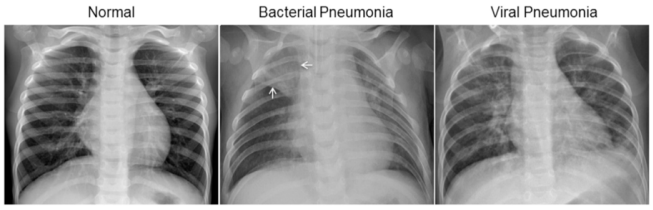


Fig. 1. Examples of Chest X-Rays [3]

pneumonia (middle) typically exhibits a focal lobar consolidation, in the right upper lobe (red rectangle), whereas viral pneumonia (right) manifests with a more diffuse interstitial pattern in both lungs.

B. Environment setup

1) Hardware: For ORB and SVM classification, an ordinary high-performance computer is enough, like 16G memory, i7(2.3 GHz) and a 256G solid-state drive (SSD) disk. However, training a deep neural network should use GPU to accelerate the process. In this report, a Google Cloud GPU is used. A virtual machine instance with four core of CPU, 16G memory and an NVIDIA Tesla K80 is used. Concerning the detail setup guide, please refer google guide and other web pages [19] [35]

2) Software: To test the ORB & SVM classification, A python program which was initially used to classify plants are ported [36]. It was modified to use the new dataset and ran it on a laptop. An iteration of the test needs about 4 hours. For transfer learning, a python program based on a python kernel in the Kaggle website was rewritten [14]. Because the CNN-based method is computing intensive, so it needs to run on a VM in Google GPU Cloud. To test the Capsule network, a python capsule network implementation that aims to detect brain tumors was ported to the pneumonia dataset [37]. It also needs to be run on the GPU VM.

C. Data Augmentation Design

In this report, three data augmentation algorithms will be evaluated. It can be seen from Table II, Aug0 means using the original dataset without augmentation. Aug1 means simple geometrical transform of the image: such as randomly flip horizontally and vertically, randomly rotates within 0.05 degrees, horizontal shear within the range 0.05 times the image width and zoom in within 0.05 times. While Aug2 is a more complicated transform than Aug1. Besides all transforms of Aug1, it also does slightly horizontal and veridical shift. To avoid the exploration of the combination of the data augmentation models and classification algorithms, this report only evaluates the effects of different augmentation algorithms on VGG16 which is the best classification algorithm for this report. However, to analyze the effects of data augmentation, all three classification algorithms will also be evaluated on Aug0 and on the best augmentation model got from this test.

TABLE II
AUGMENTATION MODELS

Augmentation Model	Augmentation parameters
Aug0	No Aug
Aug1	rotation_range=0.05, shear_range=0.05, zoom_range=0.05, horizontal_flip=True, vertical_flip=True
Aug2	rotation_range=3, width_shift_range=0.05, height_shift_range=0.05, shear_range=0.05, zoom_range=0.05, fill_mode='constant', cval=0., horizontal_flip=True, vertical_flip=True

D. ORB and SVM Application Experiments Design

The ORB, VLAD and SVM classification is chosen as the baseline. Two experiments will be conducted: First is classifying Normal and Pneumonia with the original dataset. Second does the same classification but with the best augmentation models.

E. Transfer Learning Experiments Design

Because the chest X-Ray dataset is small and different to ImageNet whose weight will be used in the transfer learning experiments, therefore three group experiments will be conducted to fine tune the final model. The first group of experiments aim to evaluate the effect of classification layer size on the final classification accuracy. Five models will be used on two CNN: VGG16 and InceptionV3, showed in Table III. In the second column, the classification model is described. For example, model3 consists of eight layers after the ConvLayers which are: Global Average Pooling (GAP) layer, FC layer with 512 neurons, Dropout layer with 50% drop rate, second FC layer with 256 neurons, second dropout layer with 50% drop rate, third FC layer with 128 neuron, third dropout layer with 50% drop rate and a classification layer with a SoftMax activation function. The third and fourth columns list the parameters needed to be trained in VGG and InceptionV3. They will be used to indicate the complexity of the model.

The second group experiment aims to evaluate how many ConvLayers should be unfrozen and be trained. A total of three experiments will be conducted (see Table IV). The first experiment will evaluate the results of the best classification model with an unfrozen ConvLayer. Because the training parameter of the last ConvLayer is quite large, to prevent overfitting, the second experiment will use a smaller classification model. For testing the limit of the number of unfrozen ConvLayers, the

TABLE III
CLASSIFICATION LAYER MODEL CONFIGURATION

	Configuration	VGG16 Training parameter	InceptionV3 Training parameter
Model1	GAP → FC(4096)→ FC(4096)→ Softmax	18,890,754	25,182,210
Model2	GAP → Softmax	1,026	4,098
Model3	GAP → FC(512) → Dropout(0.5) → FC(256) → Dropout(0.5) → FC(128) → Dropout(0.5) → Softmax	427,138	1,213,570
Model4	GAP → FC(512) → Dropout(0.5) → Softmax	263,682	1,050,114
Model5	GAP → FC(512) → Dropout(0.5) → FC(512) → Dropout(0.5) → FC(256) → Dropout(0.5) → Softmax	657,154	1,443,586

third experiment will unfreeze one more ConvLayers of the better one in the previous two models.

TABLE IV
CONFIGURATION OF FINE-TUNED CONVLAYER MODEL

	Configuration
ConvLayer Model 1	Best classification model with an unfrozen ConvLayer
ConvLayer Model 2	Smaller classification model with an unfrozen ConvLayer
ConvLayer Model 3	Better model in previous two model with two an unfrozen ConvLayer

The third group experiment will fine tune other parameters based on the best model in the previous two experiment groups, such as increasing the drop rate of the dropout layer, reducing the learning rate, adding a batch normalization layer which will make learning more stable and quicker.

F. Capsule Neural Network Design

For CapsNet, the feature map number, the size of PrimaryCaps layer and input image size will impact on the performance of the classification. Thus, Table V shows the experiments aimed at evaluating the effects of those parameters.

IV. EXPERIMENTS RESULTS

A. Data Augmentation

In Table VI, the first column is the augmentation algorithms used in the test, the second column is the total training images generated by augmentation, and the last column is the average accuracy achieved by VGG16 transfer learning with all default parameter. From the result it can be seen, the aug1 is a better augmentation model than Aug2, therefore in following experiments will use aug1 as the default augmentation model.

TABLE V
EXPERIMENTS CONFIGURATION OF CAPSNET

	Configuration
Test1	Aug0
Test2	Aug1
Test3	Aug1 with 64 feature maps and 64 input size
Test4	Aug1 with 64 feature maps and 128 input size
Test5	Aug1 with 32 feature maps and 64 input size
Test6	Aug1 with 32 feature maps and 128 input size
Test7	Aug1 with 32 feature maps and 48 input size
Test8	Aug1 with 24 feature maps and 64 input size
Test9	Aug1 with 16 feature maps and 64 input size
Test10	Aug1 with 32 feature maps, 64 input size and half primary capsule (4)
Test11	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16)
Test12	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and one fourth capsule channel (8)
Test13	Aug1 with 24 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16)
Test14	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16) with more image by augmentation (10,000)

TABLE VI
DATA AUGMENTATION EXPERIMENTS RESULT

Aug algorithms	Total training images	VGG16 Accuracy
Aug0	5,232	0.882
Aug1	5,232	0.898
Aug2	5,232	0.895
Aug1	10,000	0.902
Aug2	10,000	0.879

B. ORB and SVM Classification

TABLE VII
ORB AND SVM CLASSIFICATION EXPERIMENTS RESULTS

Augmentation	Accuracy
No Aug	0.740
Aug 20,000 images	0.776

In Table VII, the first column is the augmentation methods, and the second column is the average accuracy of the linear SVM classifier with ORB features. It can be seen the augmentation with more images will increase the accuracy.

C. Transfer Learning Classification

TABLE VIII
EXPERIMENTS RESULT OF EVALUATING CLASSIFICATION MODEL

	VGG16	InceptionV3
Model1	0.881	0.629
Model2	0.631	0.818
Model3	0.898	0.875
Model4	0.873	0.857
Model5	0.885	0.869

In Table VIII, it can be seen that VGG16 is better than InceptionV3, and Model3 in VGG16 is the best classification model. Thus, in the following experiment, VGG16 and Model3 will continue to be fine-tuned.

TABLE IX
EXPERIMENTS RESULT OF FINE-TUNED CONVLAYER

Model	VGG16
Model3 with last unfrozen ConvLayer	0.883
Model2 with last unfrozen ConvLayer	0.924
Model2 with last two unfrozen ConvLayer	0.900

From Table IX, it can be seen that the classification model2 with the last unfrozen ConvLayer was the best model in all three experiments. Thus, the following experiments will continue to fine tune it.

TABLE X
EXPERIMENTS RESULT OF FINE-TUNED OTHER PARAMETERS

Configration	VGG16
Model2 with last unfrozen ConvLayer, lr 0.0009 and lr_decay 0.8	0.900
Model2 with last unfrozen ConvLayer, lr 0.001 and lr_decay 0.5	0.873
Model2 with last unfrozen ConvLayer and 20,000 augmentation image	0.871
Model2 with last unfrozen ConvLayer lr 0.0005, lr_decay 0.5 and 10,000 augmentation image	0.902
Model3 with last unfrozen ConvLayer drop rate 0.7	0.885
Model3 with drop rate 0.7	0.906
Model3 with drop rate 0.7 and 20,000 augmentation image	0.922
Model3 with drop rate 0.7 and 30,000 augmentation image	0.922
Model2 with last unfrozen ConvLayer, batch normal layer, drop rate 0.5	0.912
Model2 with last unfrozen ConvLayer, batch normal layer, drop rate 0.5 and 20,000 augmentation image	0.906
Model2 with last unfrozen ConvLayer, batch normal layer, dropout 0.7, fc layer, dropout 0.5	0.916
Model2 with last unfrozen ConvLayer, batch normal layer, dropout 0.7, fc layer, dropout 0.5 and 20,000 augmentation image	0.875

The experiments in Table X are explorational testing. The most successful models in previous experiments: Model3 and Model2 with the last unfrozen ConvLayer are chosen as the baseline. Then according to the results and the effects of dropout, learning rate and more training data, the parameters will be adjusted. In all the experiments, Model2 with the last unfrozen ConvLayer and all other default parameters still has the best results (see Table IX and Table X).

D. Capsule Neural Network

Table XI shows the experiments processes and results by capsule network. The augmentation first is evaluated. The aug1 is better than no augmentation, and the number of feature maps, input size, number of primary capsules, number of capsule channels and number of training images vary. The best result comes from tests no.11 and 13. They both have fewer feature maps, primary capsules and capsule channels. The difference between them is the number of feature maps. One has 24 feature maps while the other has 32 feature maps.

TABLE XI
EXPERIMENTS RESULT OF CAPSNET

No	Configuration	CapsNet
1	Aug0	0.748
2	Aug1	0.788
3	Aug1 with 64 feature maps and 64 input size	0.737
4	Aug1 with 64 feature maps and 128 input size	0.627
5	Aug1 with 32 feature maps and 64 input size	0.798
6	Aug1 with 32 feature maps and 128 input size	0.784
7	Aug1 with 32 feature maps and 48 input size	0.756
8	Aug1 with 24 feature maps and 64 input size	0.798
9	Aug1 with 16 feature maps and 64 input size	0.765
10	Aug1 with 32 feature maps, 64 input size and half primary capsule (4)	0.811
11	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16)	0.825
12	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and one fourth capsule channel (8)	0.752
13	Aug1 with 24 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16)	0.825
14	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16) with more image by augmentation (10,000)	0.788

E. Verify on OCT Dataset

TABLE XII
EXPERIMENTS RESULT ON OCT DATASET

No	Model	Accuracy
1	Model2 with last ConvLayer	0.934
2	Model3	0.828
3	Inceptionv3 Model2	0.791
4	Model2 with last two unfrozen ConvLayer	0.921
5	VGG16 with last ConvLayer $\rightarrow$ 4096 FC $\rightarrow$ 0.7 dropout $\rightarrow$ 2048 FC $\rightarrow$ 0.5 dropout	0.954
6	VGG16 with last ConvLayer $\rightarrow$ 4096 FC $\rightarrow$ 0.7 dropout $\rightarrow$ 2048 FC $\rightarrow$ 0.7 dropout $\rightarrow$ 2048 FC $\rightarrow$ 0.5 dropout	0.937
7	VGG16 with last ConvLayer $\rightarrow$ 4096 FC $\rightarrow$ 0.8 dropout $\rightarrow$ 2048 FC $\rightarrow$ 0.7 dropout	0.938

To check if the findings can be used on other datasets, some experiments were conducted on the OCT dataset which was published together with the Pneumonia dataset but included 108,309 OCT images. From Table XII, it can be seen, the best result obtained from test 5.

V. DISCUSSION

A. The Effects of Data Augmentation

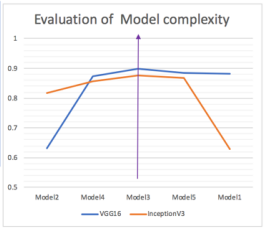
TABLE XIII
THE COMPARISON OF DATA AUGMENTATION EXPERIMENTS

Model	Accuracy without Augmentation	Accuracy with Augmentation
ORB & SVM	0.740	0.776
VGG16	0.883	0.923
INV3	0.844	0.875
Caps Net	0.774	0.856

Table XIII shows a summary of all the experiment test result between no augmentation and augmentation. It can be seen clearly that augmentation will improve the performance

Model	VGG16 Params	InceptionV3 Params	VGG16	InceptionV3
Model2	1,026	4,098	0.63086	0.81836
Model4	263,682	1,050,114	0.87305	0.85742
Model3	427,138	1,213,570	0.89844	0.875
Model5	657,154	1,443,586	0.88477	0.86914
Model1	18,890,754	25,182,210	0.88086	0.62891

a



b

Fig. 2. The Evaluation of Model Complexity

regardless of the model. That is because augmentation geometrically transform the picture which will facilitate the machine learning algorithm to learn the underground feature without the impact of rotation and scale. However, from Table VI, it can be seen that complicated transforms are not always better than simple ones. Too complicated transforms will introduce some noise in the feature that will disturb the learning process.

B. The Finding on Fine-tune of Transfer Learning

1) The Effects of Model Complicity of Neural Network: the left table of Figure 2 is a combination of Table III and Table VIII and sorted by the number of parameters in ascending order. It can be seen that the number of parameters have a significant impact on the accuracy. Too many and too few parameters will get poor results. The right graph of Figure 2 shows that the highest results of VGG16 and InceptionV3 are in model3 that has proper size of parameters that match to the size of the database.

TABLE XIV
EVALUATION OF DROPOUT, BATCH NORMALIZATION AND LEARNING RATE FOR MODEL2 WITH LAST UNFROZEN CONVLAYER

Learning rate	Decay rate	Training image	Dropout1	Dropout2	BN layer	VGG16
0.001	0.9	20000	0.5	NA	no	0.871
0.001	0.5	5323	0.5	NA	no	0.873
0.001	0.9	20000	0.7	0.5	yes	0.875
0.0009	0.8	5323	0.5	NA	no	0.900
0.0005	0.5	10000	0.5	NA	no	0.902
0.001	0.9	20000	0.7	NA	yes	0.906
0.001	0.9	5323	0.5	NA	yes	0.912
0.001	0.9	5323	0.7	0.5	yes	0.916
0.001	0.9	5323	0.5	NA	no	0.924

2) The Effects of Techniques to Preventing Overfitting: Table XIV shows the explorational test results of model2 with the last unfrozen ConvLayer. Because the whole training process tends to overfit, no single factor has a stable and significant impact on final accuracy.

When comparing the result of model3 with different conditions (see Table XV), it can be seen that the increasing dropout rate and augmentation number in each training iteration will continually increase the accuracy. The opposite is the model with the last unfrozen ConvLayer. That is understandable because the last ConvLayer has too many parameters. Therefore, the training process is overfitting.

TABLE XV
EVALUATION OF DROPOUT, BATCH NORMALIZATION AND LEARNING RATE FOR MODEL3

Model	training image	dropout1	VGG16
Model3 with last unfrozen ConvLayer	5323	0.7	0.885
Model3	5323	0.5	0.898
Model3	5323	0.7	0.906
Model3	20000	0.7	0.922
Model3	30000	0.7	0.922

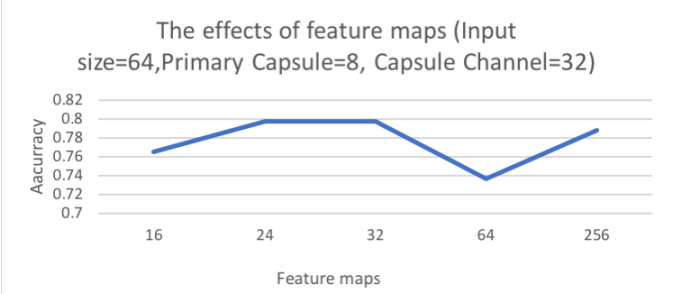


Fig. 3. The Effects of Feature Maps in CapsNet

C. The Finding on Capsule Network

1) The Effects of Feature Maps: A series of experiments can unveil the effects of feature maps through fixing the input size(64), number of primary capsules(8), number of capsule channel(32) and varying the number of feature maps. The results in Figure 3 show that the model with 24 and 32 feature maps got the best results.

2) The Effects of Input Size: A series of experiments can unveil the effects of feature maps through fixing feature map size(32), number of primary capsules(8), number of capsule channel(32) and varying the input size. Figure 4 shows that the model with input size 64 got the best accuracy.

3) The Effects of Primary Capsule: A series of experiments can unveil the effects of feature maps through fixing feature maps size(32), input size(64), number of capsule channels(32) and varying the number of primary capsules. It can be seen in Figure 5 that the model with primary capsule 4 got better accuracy than primary capsule 8.

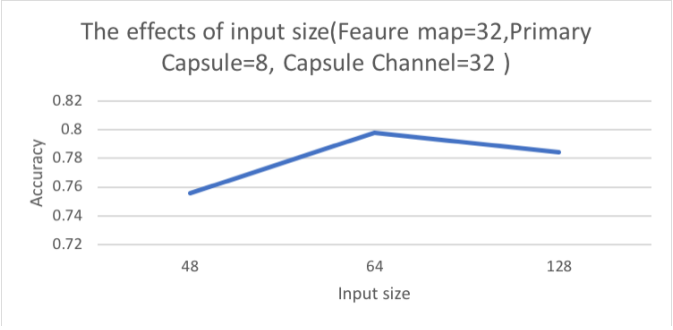


Fig. 4. The Effects of Input Size in CapsNet

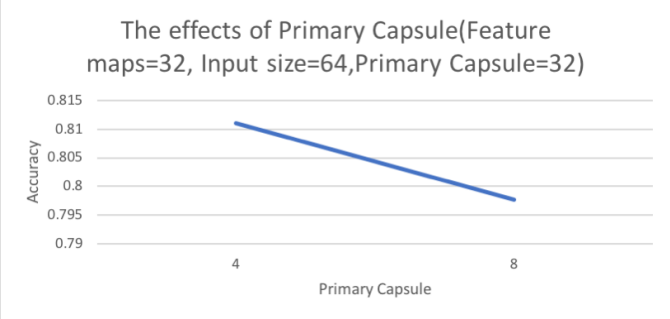


Fig. 5. The Effects of Primary Capsule in CapsNet

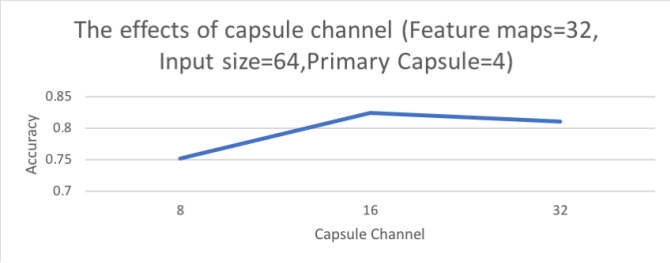


Fig. 6. The Effects of Capsule Channel

4) The Effects of Capsule Channel: A series of experiments can unveil the effects of feature maps through fixing feature maps size(32), input size(64), number of primary capsules (4) and varying the number of capsule channels. It can be seen in Figure 6 that the model with capsule channel 16 got the best accuracy.

5) The Best Model: The model with combination of feature maps size (32 or 24), input size (64), the number of primary capsule (4) and the number of capsule channels (16) should get the best results. This can be verified by the results of test 11 and 13 in Table XI : they are the best of all the tests. This also agrees with the finding in transfer learning: The complexity of a model should match the scale of a dataset.

D. Horizontal Comparison

To evaluate the performance of the models in this report, Table XIII compares the best results of different models on the same pneumonia dataset. From Table XVI, it can be seen that a neural network-based method is significantly better than the traditional method because it is a useful feature learner while the traditional method just a feature-ORB. In version 2 of the dataset, the best model, VGG16 in this report, got slightly lower accuracy and recall than the state-of-art result but got a higher specificity. On the latest dataset, the performance of VGG16 was generally higher. The VGG16 model released the last ConvLayer, so it would learn the specific features of the dataset. That should significantly help to improve the performance very much. Kermany's work also retrains the ConvLayer of InceptionV3, but the model overfits too much to get a good test performance. The reason why our model does not overfit too much maybe because the VGG16 model is not as complicated as InceptionV3.

TABLE XVI
HORIZONTAL COMPARISON OF RESULT

Model	Normal vs Pneumonia			Bacteria vs Virus		
	Accuracy	Specificity	Recall	Accuracy	Specificity	Recall
Baseline	0.776	0.809	0.776	0.643	0.640	0.585
VGG16 [34][1]	0.923	0.926	0.923	0.903	0.909	0.850
VGG16 [38][2]	0.938	0.944	0.938	0.915	0.917	0.879
Inception V3	0.869	0.854	0.869	0.851	0.860	0.779
Caps Net	0.824	0.846	0.824	0.862	0.875	0.785
State-of-art [3] [3]	0.928	0.901	0.932	0.907	0.909	0.886

[1] This result got from the version 2 of Kermany's dataset
[2] This result got from the version 3 of Kermany's dataset that is a new released by authors to fix some error in version 2 dataset
[3] The state-of-art result got from the research of Kermany et al. (2018), which is from a transfer learning based on InceptionV3

E. Finding in Verifying on OCT dataset

From Table XII, it can be seen that the best model comes from test 5 instead of test 1. The new model adds complicated FC layers, therefore, the whole complicity is better matched with the new dataset. The unfrozen two ConvLayers will make the system too complicated for the new dataset and therefore cannot find the local maxima. The best result is slightly lower than the start-of-art result of Kermany's work (96.6%) However, this experiment result also confirms our finding. The specific feature is most important to improve accuracy. The proper model complexity will help to find the best result.

F. Answerers of research questions

After the discussion of all experiment result, the research questions can be answered as follows:

How can data mining, including traditional ways and CNN-based classification be used to classify small medical images dataset?

From the above discussion, it can be seen that traditional classification and CNN-based classification can also be used in medical image classification. However, the traditional classification algorithm needs to spend many efforts on feature engineering, while the CNN-based classification algorithms can learn feature automatically. Moreover, the CNN-based transfer learning classification can utilize the powerful feature extraction capability trained on massive image dataset to get the significant result on the small dataset. On the other hands, training a capsule neural network from scratch on the small dataset also can get a good result.

How do the performances of different data mining methods compare with each other?

In all the methods, the CNN based transfer learning is the best method. The second-best model is the Capsule network. The last is the ORB & SVM classifier. CNN based methods are better than traditional methods because they can learn features automatically and effectively. The reason why transfer learning CNN-based classification gets the best result is their

ConvLayers are trained from a massive dataset. Therefore, they can learn comprehensive primary features that even cannot learn from the new dataset, and then the high-level features are learned from the new dataset. The best result in this report is slightly higher than the state-of-art result on the same pneumonia dataset by 1% on the general accuracy.

G. Conclusions and Future work

Due to the importance of medical image classification and the particular challenge of the medical image-small dataset, this report chose to study how to apply CNN-base classification to small chest X-Ray dataset and evaluate their performance. From the experiments, the following finding can be presented. The CNN based transfer learning is the best method of all three methods. Capsule network is better than ORB & SVM classifier. Generally speaking, CNN based methods are better than traditional methods because they can learn and select features automatically and effectively; The best results come from the transfer learning of VGG16 with one retrained ConvLayer, which is slightly higher than the start-of-art result. With the unfrozen ConvLayer, the specific feature can learn from the new dataset. Therefore, the specific feature is the most important factor to improve accuracy; The balance of a model's power of expression and overfitting is essential. A too simple network usually cannot learn enough from the data, and therefore cannot get high accuracy. On the other hand, a very complex network is hard to train and tends to overfit quickly. As a result, the accuracy is still low. Only a network model with proper size and other effective methods preventing overfit, such as proper dropout rate and proper data augmentation, can get the best results.

However, because of the limited time, future research needs to be done: In transfer learning, training a fine tuned deep neural network with unfrozen ConvLayers tends to overfit. What effective methods can be done to stabilize the training process? Other more powerful CNN model, such as ResNetv2 and ensemble of multiple CNN models have not been evaluated, but they could possibly improve the results; Visualization needs to be added to improve the understanding and explanation of the results of the CNN-based system, because those are essential for the adoption of a CNN-based system in real clinical applications.

Acknowledgment

I would first like to thank my thesis advisor Dr. Shafiq Alam of Whitireia New Zealand. The door to Dr. Alam's office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this report to be my own work but steered me in the right the direction whenever he thought I needed it.

Moreover, I would also like to acknowledge Dr. Zawar Shah of Whitireia New Zealand. He gave me many suggestions on how to organize my report and pointed out my errors, carefully pushing me to improve my writing skills. Moreover, without his reminders regarding key points of this research schedule, it would have been hard for me to finish this research on time.

References

[1] World Health Organization, "Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children," 2001.

[2] CDC, "Pneumonia Can Be Prevented—Vaccines Can Help," 2017.

[3] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[4] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[5] A. Jiménez-Sánchez, S. Albarqouni, and D. Mateus, "Capsule networks against medical imaging data challenges," *arXiv preprint arXiv:1807.07559*, 2018.

[6] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9049–9058, 2018.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, Ieee, 2009.

[8] Stanford Vision Lab, "ImageNet Summary and Statistics (updated on April 30, 2010)," 2010.

[9] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for sar target recognition," *IEEE Geoscience and remote sensing letters*, vol. 13, no. 3, pp. 364–368, 2016.

[10] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *arXiv preprint arXiv:1803.01229*, 2018.

[11] C. N. Vasconcelos and B. N. Vasconcelos, "Increasing deep learning melanoma classification by classical and expert knowledge based image transforms," *CoRR, abs/1702.07025*, vol. 1, 2017.

[12] J. Zhou, Z. Li, W. Zhi, B. Liang, D. Moses, and L. Dawes, "Using convolutional neural networks and transfer learning for bone age classification," in *Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference on*, pp. 1–6, IEEE, 2017.

[13] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[14] Rajat Goel, "Prediciting Pneumonia with the help of transfer Learning," May 2018.

[15] S. Hoo-Chang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, p. 1285, 2016.

[16] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, pp. 3856–3866, 2017.

[17] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, vol. 2, 2010.

[18] P. Afshar, A. Mohammadi, and K. N. Plataniotis, "Brain tumor type classification via capsule networks," *arXiv preprint arXiv:1802.10200*, 2018.

[19] Google, "Google gpu cloud," 2018.

[20] M. Sonka, "Medical image processing and analysis," *Handbook of Medical Imaging*, vol. 2, 2000.

[21] R. Paredes, D. Keysers, T. M. Lehmann, B. Wein, H. Ney, and E. Vidal, "Classification of medical images using local representations," in *Bildverarbeitung für die Medizin 2002*, pp. 171–174, Springer, 2002.

[22] N. Parveen and M. M. Sathik, "Detection of pneumonia in chest x-ray images," *Journal of X-ray science and technology*, vol. 19, no. 4, pp. 423–428, 2011.

[23] J. C. Caicedo, A. Cruz, and F. A. Gonzalez, "Histopathology image classification using bag of features and kernel functions," in *Conference on Artificial Intelligence in Medicine in Europe*, pp. 126–135, Springer, 2009.

[24] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE international conference on*, pp. 2564–2571, IEEE, 2011.

[25] A. Mueen, S. Baba, and R. Zainuddin, "Multilevel feature extraction and x-ray image classification," *Journal of Applied Sciences*, vol. 7, no. 8, pp. 1224–1229, 2007.

[26] X. Yuan, Z. Yang, G. Zouridakis, and N. Mullani, "Svm-based texture classification and application to early melanoma detection," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pp. 4775–4778, IEEE, 2006.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[28] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*, pp. 844–848, IEEE, 2014.

[29] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 3462–3471, IEEE, 2017.

[30] V. P. Vianna, "Study and development of a computer-aided diagnosis system for classification of chest x-ray images using convolutional neural networks pre-trained for imagenet and data augmentation," *arXiv preprint arXiv:1806.00839*, 2018.

[31] T. Iesmantas and R. Alzbutas, "Convolutional capsule network for classification of breast cancer histology images," in *International Conference Image Analysis and Recognition*, pp. 853–860, Springer, 2018.

[32] F. Beşer, M. A. Kizrak, B. Bolat, and T. Yildirim, "Recognition of sign language using capsule networks," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2018.

[33] E. Xi, S. Bing, and Y. Jin, "Capsule network performance on complex data," *arXiv preprint arXiv:1712.03480*, 2017.

[34] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification(V2)," *Mendeley Data*, vol. v2, 2018. itemType: dataset.

[35] James Lee, "Setting Up a Google Cloud Instance GPU for fast.ai for Free," Dec. 2017.

[36] Hernaldo Jesús Henríquez Núñez, "Python and OpenCV code for object classification using images," 2016.

[37] Spencer Chang, "Tumor-CapsNet," 2018.

[38] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification(V3)," *Mendeley Data*, vol. v3, 2018. itemType: dataset.