**Automatic Classification of**

**Medical Modality using**

**Convolutional Neural Network**

Moana Yamanouchi (27742407)

Bryan Soh (27031055)

Supervised by: Dr. Mohammad Reza

Literature Review

FIT3161 Computer Science Project 1

Monash University Malaysia

School of Information Technology

**Abstract**

**With rapid technological advances in modalities, new algorithms and methods are required in order to differentiate the characteristics of the medical images within the large image collection. This task can be achieved using image classification which assigns multitude of input image features from dataset into a specific category through extraction and fusion. The two main key challenges for medical image classifications are 1) identifying the differences ranging from subtle to distinct among modalities and 2) the dataset provided in medical dataset is imbalanced and limited labelled dataset which will increase the risk of overfitting. In this paper, we will analyse various research papers on image modality classification using convolutional neural network (CNN) and fusion techniques. We will then hypothesize which method is suitable for our topic based on its accuracy level.**

**Keywords: Image classification, medical modality, convolutional neural network, fusion technique.**

Contents

[1. Introduction 3](#_Toc513813032)

[1.1 History 3](#_Toc513813033)

[1.2 Overview 4](#_Toc513813034)

[1.3 Problem Statement 4](#_Toc513813035)

[1.4 Motivation 4](#_Toc513813036)

[1.5 ImageClef 2016 Database 4](#_Toc513813037)

[1.6 Feature Extraction 5](#_Toc513813038)

[2. Convolutional Neural Network 5](#_Toc513813039)

[2.1 Architecture 6](#_Toc513813040)

[2.2 Feature Extraction 6](#_Toc513813041)

[3. Fusion Technique 7](#_Toc513813042)

[3.1 Early Fusion 7](#_Toc513813043)

[3.2 Late Fusion 7](#_Toc513813044)

[4. Related Works 7](#_Toc513813045)

[4.1Types of CNN Architecture 7](#_Toc513813046)

[4.2 Analysis on ImageClef 2016 7](#_Toc513813047)

[4.3 Overcome Overfitting 8](#_Toc513813048)

[4.4 Classification Accuracy 9](#_Toc513813049)

[5. Conclusion 10](#_Toc513813050)

[References 10](#_Toc513813051)

# Introduction

Medical imaging is an essential process for diagnosis, monitoring and treatment to make accurate medical decisions. There are various types of medical imaging such as Computer Tomography (CT), Magnetic Resonance Imaging (MRI), X-ray and Ultrasound. Using images for comparative studies is an essential technique for diagnostic decision-making for health professionals based on their knowledge and analysing from a patient’s data. In recent years, improvements in healthcare have been made due to the technological advancement of medical modality. The growing focus of development in medical devices and high demand for both medical and business field creates many high investment opportunities in the medical sector.

The most commonly used database for storing and accessing medical images is text-based Picture Archiving and Communication Systems (PACS). Although PACS search is useful for health professionals to search known identifiers and characteristics, it cannot retrieve information based on visual features[1]. Furthermore, PACS image retrieval becomes infeasible for large imaging datasets which is inevitable in modern healthcare as diverse ranges of imaging data is required and is compounded with development of technological innovations in medical imaging modality. With such large volume, manually labelling images become uneconomical and prone to human error.

To eliminate these limitations, content-based image retrieval (CBIR) was introduced[2] by assessing quantitative measures and objective features followed by extracting similar information from a feature representation at a pixel-level. This image search technique utilizes visual features which includes mainly colour, shape, and texture for medical images. The two main key components of CBIR are providing 1) suitable feature representation and 2) effective similarity measures. The main challenge for CBIR is the semantic gap between low level images and high level concept; machine learning has become a prominent technique to reduce this gap. Identifying an accurate feature representation and specifically classifying the images is an important technique in CBIR. This task can be achieved by using medical modality classification.

## 1.1 History

Convolutional Neural Networks are used in deep learning and in a plethora of computer vision tasks today. However, there is some history to how its architecture and layers work the way they do. In 1962, two researchers, Hubel and Weisel[3] were experimenting to find out to what stimulus, neurons respond to. They recorded from neurons in the visual cortex of a cat, as they moved a bright line across its retina. They had discovered that only certain neurons were fired (occurence of an all-or-none action potential) only when light was on a particular position of the retina. The activity of the neurons that were activated, changed depending on the orientation of the line of light. Sometimes the neurons fired only when the line of light was moving in a particular direction. It was then discovered that neurons in the first few layers of a visual cortex will only pick up very simple features, like the orientation. This provided an overview of how the visual system operates. It constructs complex patterns in layers of understanding from simple features to more complex ones.

In 1982, a neuroscientist named David Marr conceptualized how images are reconstructed into 3D images in the visual cortex[4]. The first stage which he called *Primal Sketches,* which involved the use of optical flow and segmentation, to find the rough edges of objects in the image field. Followed by the second stage, *2.5D*, which is 2D with texture, depth and shade to provide further detail to the image. And the final stage, *3D*, reconstructing the entire scene.

Furthermore, in 1980, Kunihiko Fukushima had proposed and created a neural network for visual pattern recognition which could learn without any help and acquired an ability to recognize distinct patterns in an image based on geometric similarity of their shapes without being influenced by their positions[J]. The network was nicknamed “Neocognitron”. After the experiment was fully carried out, the self-organized network had a structure which was very similar to the hierarchy of the visual system that had been proposed by Hubel and Weisel. With all this groundbreaking research, there has been several biologically inspired CNN methodologies that have been proposed throughout the recent years, each with their own unique tweaks and incorporated techniques.

## 1.2 Overview

With rapid technological advances in modalities, new algorithms and methods are required in order to differentiate the characteristics of the medical images within the large image collection. This task can be achieved using image classification which assigns multitude of input image features from dataset into a specific category through extraction and fusion.

## 1.3 Problem Statement

A challenge for modality classification is identifying the differences ranging from subtle to distinct among modalities, along with the variation in the appearance of individual diseases which affects the application of classification techniques. For example, the distinct difference between X-ray and CT scan must be classified into two different modality classes. Whereas, the subtle differences between fractured and dislocated leg X-ray images needs to be classified into one modality class.

Another challenge is the imbalanced and limited labelled dataset. For example, in ImageClef 2016, training dataset of 6776 labelled medical images were provided, where approximately 42% of datasets were statistical figures, graphs, charts (GFIG). Whereas, ImageNet 2012 [5] consists of 1.2 million images. Statistically comparing these two datasets, if both accuracy resulted in same percentage, ImageNet will still obtain higher accuracy due to the difference in the size. The imbalanced and limited labelled data will lead to risk of overfitting which will be further discussed later on.

## 1.4 Motivation

Various research papers on image modality classification using convolutional neural network (CNN) and fusion techniques will be analyzed. The main goal is to accurately and efficiently classify the images regardless of the differences in visual characteristics.

The main objectives are:

* To critically summarize and evaluate previous research papers
* Explain the improvements of image modality classification in recent years
* To identify different types of CNN architecture and analyse its differences
* To identify different methods of fusion technique in CNN and analysis its advantages and disadvantages
* To include arguments/ideas while maintaining logical progression

## 1.5 ImageClef 2016 Database

For comparative analysis purpose based on accuracy result, information from ImageClef 2016 will be used in Section 4.2 and 4.4. Medical task in ImageClef is an ongoing evaluation campaign that is a part of the Conference and Labs of the Evaluation Forum (CLEF) initiated since 2004. Five subtasks were provided in 2016 are:

* Compound Figure Detection
* Multi-label classification
* Figure separation
* Subfigure Classification
* Caption Prediction

Subfigure classification is, as shown above, is one of the subtasks which was already introduced to a similar subtask in 2011. The dataset used in modality classification are figures from multi-label classification task that is then distributed into labelled subfigures.

The aim is to classify 30 heterogeneous classes ranging from various diagnosis images to generic biomedical illustrations.

## 1.6 Feature Extraction

The two main feature extractions are 1) handcrafted features and 2) automatic feature extraction using deep learning. Handcraftedfeatures for modality image classification have been proposed such as intensity histogram, local binary pattern (LBP) and scale-invariant feature transform (SIFT) which are then are encoded into bag-of-words (BoW).

For automatic feature extraction, amongst various deep learning techniques, convolutional neural network (CNN) has become a popular technique as it outperformed other conventional methods based on its high accuracy result and low test error rate [6].

# Convolutional Neural Network

CNN is a deep learning model which is a type of artificial neural network that analyses data using mathematical model. It is widely used in medical image analysis as it simplifies the training process by implemented trained end-to-end using supervised learning algorithm.

Deep learning is a subset of Artificial Intelligence (AI) which constructs a machine learning model that primarily focuses on using appropriate hierarchical representation of the data to draw conclusions. It is an architecture that is used in various machine learning application especially for a more detail-oriented and accurate image classification. Artificial neural network (ANN) is a class of machine learning algorithm that was biologically inspired by the mechanism of the human brain. In this section, the fundamental concept of feed-forward neural network will be introduced as a basis of deep model.

In a supervised learning algorithm, it consists of training dataset *D ={( xi, yi)}* which consists of images *xi* associated with labelled data *yi:*  Input feature *x* produces a predicted score *ŷ* using a score function *ŷ* = f(xi, W)where matrix weight *W* is a parameter. The loss function *l(ŷ, y)* is then applied to compare with the actual target labelled data *y* and *ŷ* to compute its disagreement. The gradient of the loss function is then computed where *W* is updated using backpropagation by applying recursive application of chain rule by computing the gradient (partial derivative) *f* on *x.* Each neurons, therefore has a value of *f* which is then adjusted during backpropagation by multiplying the local gradient and its input value. An example would be classification, regression, object detection, semantic segmentation and image captioning.

In an unsupervised learning algorithm, it is trained to learn an underlying hidden structure of the data without using labelled data *y.* An example of unsupervised learning is clustering, dimensionality reduction, feature learning and density estimation. A common technique that is used to evaluate the quality is by applying feature extractor on a supervised dataset.

Training CNN using dataset for supervised learning algorithm will have space complexity *O(N2)* as it contains both *N* number of image *x* and *N* number of labelled data *y* in the dataset. Whereas, unsupervised learning algorithm will have space complexity of *O(N)* as only image *x* is required in the dataset. Therefore, based on the comparison of space complexity, unsupervised learning algorithm will be more efficient as it requires less memory space.

## 2.1 Architecture

CNN consists of three main layers which are convolutional layer, pooling layer and fully-connected layer. It views inputs as images which is what distinguishes from other neural network. This assumption is taken as an advantage and structures the neurons in three-dimensional volume which significantly reduces the total number of parameters in the network.

*Convolutional Layer*

In convolutional layer, it consist a filter (or kernel) of array weight *W* which slides/convolves over the height and width of the input value using a dot product to produce two dimensional activation map. The filter is slide across the image depending on the stride length and padding. This map detects the specified visual feature. The network will initially learn low level features and as it progresses in subsequent layers, a more complex abstract feature becomes the input.

Parameter sharing scheme is implemented to control number of parameters which constrains neurons in each depth slice. This is due to the sparse connection between layers, causing the area of the filter to decrease. By constraining the weights, it will increase the algorithm efficiency and by using constant weights, it will significantly reduce its memory storage.

*Pooling Layer*

The main function of pooling layer is to downsample the spatial size of the image and number of parameters. The most commonly used operation is max-pooling which extracts the highest input value within the targeted filter to compile the highest activations.

*Fully Connected Layer*

The fully connected layer is the final layer of the network where, unlike pooling layer, each neuron in the input volume is fully connected to the next layer.

## 2.2 Feature Extraction

There are twomain methods for feature extraction using CNN architecture[7]: 1) Global feature extractor obtains a holistic representation of the image by extracting the activation from fully-connected layers. 2) Local feature extractor focuses more on a specific area which is achieved by removing the fully-connected layers and accepts activation maps as local features. The features are then combined into single vector using encoding method such as VLAD encoding, BoW encoding or Fisher Vector encoding. Sum-pooling or max-pooling is also used as local feature aggregation for its efficient computation and non-parametric measures.

# Fusion Technique

Fusion technique is a strategy to fusing different information to improve accuracy and robustness.

## 3.1 Early Fusion

Early fusion, also known as feature fusion, is a technique where all distinct features vectors are concatenated into a single feature vector. By combining these high dimensional feature vectors, it stores the effective discriminate information while simultaneously eliminates redundant information. Therefore, the total vector size would be the sum of the features dimensions due to the merging of all feature vectors. Although, this gives good performance, feature discrimination is neglected as it completely relies on the learning algorithm. Therefore, sum (or average) pooling or maximum pooling is implemented to extract only the significant feature values.

## 3.2 Late Fusion

Late fusion integrates separate classification scores. This method is achieved by using features that extracted from individual channel and feeding into a classifier to extract each individual classification score. The two disadvantages in late fusion is 1) number of classifiers is same as total number of features which is computationally infeasible and 2) connection between individual features do not reflect on the classification score due to the individual complementary cues. Regardless of these disadvantages, late fusion is still used for its accurate result.

# Related Works

## 4.1Types of CNN Architecture

*AlexNet[8]*,introduced by Krizhevsky *et al.*, was the first large scale CNN achieved in image classification which outperformed all previous non-deep learning application in ILSVRC challenge in 2012 with top 5 error rate of 16% compared to 26%. In the year 2014, Simonyan and Zisserman introduced *VGGNet* which reinforced that the depth of the network correlates with the overall performance of hierarchical representation of the data.

In ILSVRC 2014, *GoogLeNet*[9] was the winner due to its high computational efficiency. The main contribution was the development of “Inception Model” by creating an efficient local network topology. The architecture uses parallel operations containing multiple receptive field sizes for convolution and pooling operation which then concatenates into single vector in the next layer. However, the total number of convolutional operation (which includes spatial location for each filer map and dot product operation) and the pooling operation would lead to high computational complexity. This lead to the improvement of the architecture where linear filter was replaced with “micro networks”, a bottleneck layer that uses 1x1 convolutions that was investigated by Lin *et al.*[10]. By implementing the extra layer, it preserves spatial dimension while reducing its depth.

Compared to *AlexNet* both deeper CNNs, *VGGNet*[11] and *GoogLeNet*, has decreased its parameter, because smaller filters were applied while sustaining the effectiveness of the receptive field. However, this increases the total space complexity, because each node requires memory allocation for forward and backpropagation.

## 4.2 Analysis on ImageClef 2016

Biomedical Computer Science Group (BCSG) obtained the highest accuracy result of 88.43% by applying early fusion technique of four different classifiers: 1) applying early fusion of both visual and textual feature which were fed into to SVM classifier with RBF kernel, 2) transfer learning was applied where features were extracted from1) and fed into pre-trained *ResNet-152[12]*, 3) adopting global visual feature by using transfer learning on *ResNet-152[12]*and extracting its features from fully-connected layer. Furthermore, transition from higher to lower dimensional feature was applied using *Projection Learning Rule* and 4) applying matrix factorization technique called non-negative matrix factorization (NMF).

Early fusion of both textual and visual features has improved the accuracy of the result. By applying the current state-of-the-art CNN architecture, *ResNet-152[12]*, has improved its accuracy result. For further improvement, fine-tuning can be applied instead of transfer learning, because transfer learned features are reflective of the original dataset (which are usually images that do not link with medical modality images) which may not reflect on the subtle differences of medical images.

NovaSearch implemented CNN to evaluate its classification performance with such unbalance and limited dataset provided by ImageClef 2016. VGG-like model (inspired by *VGG[11]*) with dropout technique had the highest result out of all the runs with 65.31% accuracy[13]. Although, it has proven that CNN is an essential tool for image modality classification based on the accuracy percentage, overcoming the challenge of overfitting was not accomplished based on their result. In the pre-processing stage, only one channel was used (output activation volume with depth (or channel) of 1) which excludes RGB images and will only view the images in a greyscale value. This may hinge the overall classification result as some modality images contain coloured features. Also by not using the depth dimension, it loses the purpose of using the CNN architecture.

## 4.3 Overcome Overfitting

Supervised learning algorithm generally requires large number of training dataset for an accurate classification result; however, acquiring such volume for medical images is difficult as it must be manually annotated by a professional which is cost and time inefficient. Therefore, class imbalance and limited dataset provided will cause overfitting due to poor generalization of the training dataset. To overcome this challenge, various techniques are applied.

*Data Augmentation*

Data augmentation is a common technique[14] used by artificially enlarging dataset. The traditional technique used involves cropping, flipping, distorting and adding noises to each image.

*Transfer Learning*

Transfer learning is a process which CNN is pre-trained on a labelled natural image dataset and is used to extract its features on other datasets. The two main strategies used are 1) pre-trained CNN is used as features extractors and 2) fine-tuning CNN [15].

Kumar *et al.[16]*compared two methods using fined-tuned CNN architecture, *AlexNet* and *GoogLeNet* 1) as an image feature extractor and apply early fusion technique by concatenating with the independent feature vector that is then uses one-to-one multiclass SVM, and 2) as a classifier using softmax probability. By fusing two different CNN architecture, it has shown that shallower architectures, like AlexNet, has the ability to learn generalized features, whereas deeper architectures, *GoogLeNet*, learn semantically relevant features.

Tajbakhsh *et al. [15]* has investigated the correlation between the depth of fine-tuning and accuracy of image classifiers where various level of tuning was tested, from shallow to deep, based on fine-tuning the number of convolutional layer. As a result, it has shown that the deeper fine-tuned CNN architecture had high accuracy rate.

*Increase Dataset*

Another simple common technique used to enrich dataset is to source additional image by extracting previous dataset[17]. For example, some teams[18, 19] from ImageClef 2016 has concatenated with ImageClef 2013 dataset from modality image classification.

One of the main challenges in medical imaging domain was insufficient amount of labelled data. This problem can be omitted by the use of unsupervised learning algorithm. An approach that is expected to have high impact in medical imaging is using Generative Adversarial Networks (GANs)[20] for synthetic data augmentation. GANs are an implicit density estimation that can learn generative model based on a game theory. The adversarial modelling framework consists of two deep CNNs, generator G and discriminator D. The goal is to optimize G and improve its ability to synthesize images until D cannot distinguish between real and fake images.

Frid-Adar *et al.*[21] implemented a proposed model known as deep convolutional GAN (DCGAN) where both generator and discriminator architecture is modelled using deep CNN, to improve the overall performance of CNN for medical image classification.

DCGAN was used to create synthesized data of liver lesions on CT images with classification performance of 85.7% sensitivity and 92.4% specificity whereas classic data augmentation yields 78.6%.

If GAN is to be used in image modality classification to synthesize modality images, its architecture cannot use lesion region of interest (ROI) as automatic classification of modalities also requires discriminating distinct differences.

## 4.4 Classification Accuracy

The ensemble method by Kumar *et al.[16]* resulted in top-1 accuracy of 82.48%.The only image classification task that was applied in this paper was colonoscopy frame classification where is classifies either *informative* or *non-informative.* Therefore, this method cannot prove that this method will work on image modality classification. However, further studies[22]have tested on medical modality classification using ensemble method along with improving image pre-processing by applying cropping on images on deeper CNN architecture with a top-1 accuracy of 85.55% for *Inception-ResNet-V2* and 85.33% for the *Inception-v4*network. Although, this paper has proven a higher accuracy result compared to BCSG at ImageClef 2016, adjusting pre-processing stage can be time and cost inefficient as it requires manually cropping and resizing each images.

Akilan *et al.[23]*have shown that late fusion approach has enhanced the classification accuracy by fusing high level features from three multi-deep pre-trained CNN architectures, *AlexNet, VGG-16,* and *Inception-v3*, which were then used as features extractors. The feature fusion was applied by encoding high to low dimensional CNN features by reducing the rectified feature dimensionality. Then feature fusion was applied using various method of pooling function to compare its accuracy.

# Conclusion

From various paper reviewed in this paper, it is evident that implementing CNN architecture is a prominent technique for image modality classification. Based on our analogy, applying GoogLeNet is a suitable architecture because of 1x1 convolution layer. For example, when extracting feature of a small region (in order to detect the subtle differences between features), 1x1 convolution layer will preserve the image size while reducing the depth. Additionally, we believe that using fine-tuning method on the architecture may improve the overall classification accuracy.

We hypothesize that implementing the early fusion technique from Kumar *et al.* [16] by extracting feature dimension through different CNN architectures and using early fusion technique is suitable for image modality classification to be able to discriminate between distinct and subtle differences between image modalities.

# References

1. Kumar, A., et al., *Content-Based Medical Image Retrieval: A Survey of Applications to Multidimensional and Multimodality Data.* Journal of Digital Imaging, 2013. **26**(6): p. 1025-1039.

2. Chang, N.-S. and K.-S. Fu, *Query-by-pictorial-example.* IEEE Transactions on Software Engineering, 1980(6): p. 519-524.

3. Hubel, D.H. and T.N. Wiesel, *Receptive fields of single neurones in the cat's striate cortex.* The Journal of physiology, 1959. **148**(3): p. 574-591.

4. Marr, D., *A theory of cerebellar cortex.* The Journal of physiology, 1969. **202**(2): p. 437-470.

5. Russakovsky, O., et al., *ImageNet Large Scale Visual Recognition Challenge.* International Journal of Computer Vision, 2015. **115**(3): p. 211-252.

6. Khan, S. and S.-P. Yong. *A comparison of deep learning and hand crafted features in medical image modality classification*. in *Computer and Information Sciences (ICCOINS), 2016 3rd International Conference on*. 2016. IEEE.

7. Ergun, H., et al., *Early and Late Level Fusion of Deep Convolutional Neural Networks for Visual Concept Recognition.* International Journal of Semantic Computing, 2016. **10**(03): p. 379-397.

8. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.

9. Szegedy, C., et al. *Going deeper with convolutions*. 2015. Cvpr.

10. Lin, M., Q. Chen, and S. Yan, *Network in network.* arXiv preprint arXiv:1312.4400, 2013.

11. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.

12. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

13. Semedo, D. and J. Magalhães. *NovaSearch at ImageCLEFmed 2016 Subfigure Classification Task*. in *CLEF (Working Notes)*. 2016.

14. Kumar, A., et al. *Subfigure and Multi-Label Classification using a Fine-Tuned Convolutional Neural Network*. in *CLEF (Working Notes)*. 2016.

15. Tajbakhsh, N., et al., *Convolutional neural networks for medical image analysis: Full training or fine tuning?* IEEE transactions on medical imaging, 2016. **35**(5): p. 1299-1312.

16. Kumar, A., et al., *An ensemble of fine-tuned convolutional neural networks for medical image classification.* IEEE journal of biomedical and health informatics, 2017. **21**(1): p. 31-40.

17. Seco, G., et al., *Overview of the ImageCLEF 2016 medical task.* Working Notes of CLEF, 2016.

18. Koitka, S. and C.M. Friedrich. *Traditional Feature Engineering and Deep Learning Approaches at Medical Classification Task of ImageCLEF 2016*. in *CLEF (Working Notes)*. 2016.

19. Valavanis, L., S. Stathopoulos, and T. Kalamboukis. *IPL at CLEF 2016 Medical Task*. in *CLEF (Working Notes)*. 2016.

20. Goodfellow, I., et al. *Generative adversarial nets*. in *Advances in neural information processing systems*. 2014.

21. Frid-Adar, M., et al., *GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification.* arXiv preprint arXiv:1803.01229, 2018.

22. Koitka, S. and C.M. Friedrich. *Optimized Convolutional Neural Network Ensembles for Medical Subfigure Classification*. in *International Conference of the Cross-Language Evaluation Forum for European Languages*. 2017. Springer.

23. Akilan, T., et al. *A late fusion approach for harnessing multi-cnn model high-level features*. in *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*. 2017. IEEE.