Otto-von-Guericke University Magdeburg

Faculty of Computer Science

Master Thesis

Classification of Clinically Significant Prostate Cancer with Multiparametric MRI

Author:

Wai Po Kevin Teng

Matriculation no.: 221219

January 04, 2022

Advisors:

Prof.Dr.-Ing. Sebastian Stober
Dr.-Ing. Marko Rak
M.Sc. Oleksii Bashkanov

Teng, Wai Po Kevin:
Classification of Clinically Significant Prostate Cancer with Multiparametric MRI
Master Thesis, Otto-von-Guericke University Magdeburg, 2022.

Abstract

Prostate cancer is one of the most reported cancer cases among the male population with high mortality rate. While prostate diseases could be identified through the elevation of prostate elevation agent (PSA), it could not justify the type of prostate diseases. Patients with elevated PSA level would be introduced to multi-parametric MRI (mpMRI) screening, where mpMRI attempts to display lesion size and its corresponding position. Patients with suspicious lesion would need to undergo biopsies to determined the tumour's malignancy degree through Gleason score assignment. However, false positive of biopsies would lead to over-diagnosis, causing over-burden on patients. This thesis attempts to leverage false positive of biopsies through deep metric learning. Unlike conventional classification task that attempts to learn a mapping from features to label, deep metric learning enforced discriminative features such that objects with similar label are close by, whereas, objects with different label are far apart from each other.

Prior to this work, we received mpMRI from Altakinik with image sequences of T2, DWI and ADC from 1873 patients. This thesis proposed state-of-the-art deep metric learning method, namely margin-based softmax loss, which is an extension of the conventional softmax loss function. Specifically, we adapted ArcFace loss in this work to learn discriminative feature embedding for better classification capability. Instead of optimising a Euclidean distance in the original softmax loss, ArcFace loss projects the embedding to a hyperphere, where geodesic distance is optimised through additive angular margin. In the ablation studies, we compared ArcFace loss with softmax loss, in the attempt to draw a decision boundary between patients with clinical significant prostate tumour malignancy. In addition to single head settings, where we attempt to optimise an embedding base on one objective function, we introduced multi-head settings, where we attempt to optimise an embedding base on several objective functions. While the primary goal remains to detect clinical significant prostate cancer patients, multi-head settings provide more constraints on the feature rich embedding. Other than utilising T2, DWI, ADC image sequences respectively as model inputs, we adapted T2+DWI+ADC as channel inputs to enrich input data for our model. Furthermore, to tackle the dilemma resulting from the alignments of 3 image sequences, we implemented late fusion, where we attempt to optimise an embedding from the concatenated embeddings of T2, DWI and ADC respectively.

The backbone model implemented in this thesis is anisotropic hybrid ResNet (AH-ResNet) to tackle low depth to width/height ratio of volumetric medical images. We

were able to demonstrate superior performance of ArcFace loss with T2+DWI+ADC as input channel in single head setting with AUC score of 0.79, accuracy of 0.73, R@1 of 0.65 and MAP@10 of 0.51, as well as, multi-head setting with AUC score of 0.79, accuracy of 0.71, R@1 of 0.65 and MAP@10 of 0.55. In this thesis, we demonstrated the capability of ArcFace loss for better discriminative feature learning, as opposed to conventional softmax loss. In addition to classification and retrieval tasks, we also presented content-based image retrieval (CBIR) in our work, where we demonstrated the credibility of visual similarity search through the ranking of embedding similarities.

Acknowledgments

"My candle burns at both ends it will not last the night but arh my friends and oh my foes it gives a lovely light"

- Roald Dahl

I would like to show my utmost gratitude to my advisors, Prof.Dr.-Ing. Sebastian Stober and Dr.-Ing. Marko Rak for their insightful inputs during thesis discussion. I am particularly in debt to my supervisor, M.Sc. Oleksii Bashkanov for his invaluable advice and professional help in structuring this thesis. I am really grateful for Virtual and Augmented Reality Lab, Faculty of Computer Science, Otto von Guericke University for allocating to me multiple GPUs as computational resources during the process of my thesis.

My gratitude goes to my flatmates, Tobias Wegmann and Sören Buckstöver for aiding me with life in Germany during my master thesis writing.

Last but not the least, I would like to dedicate this thesis for the unconditional support from my beloved partner, Chang Yan Tay and my family members. The thesis would not have been done without them.

Contents

1. Introduction

1.1 Motivation

Prostate is one of the vital organs in the male reproductive system that secretes viscous fluid which acts as a protection barrier to protect the sperm cells (InformedHealth.org [2011]). A prostate is positioned below the urinary bladder and surrounded by urethra. While the primary function of the prostate is engaged during reproduction activities, it is still susceptible to prostatic related diseases. Due to the positioning of prostate in the male body, inflammation or enlargement of prostate gland would inevitably lead to discomfort of its surrounding muscular organs, ultimately leading to long term disorder or fatality at worst. According to Xia et al. [2012], some common prostate diseases include benign prostatic hyperplasia (BPH), prostatitis and prostate cancer.

BPH is the enlargement of prostate, common among men in their middle adulthood (Roehrborn [2005]). Prostatitis, on the other hand, refers to the inflammation and swelling of prostate gland (Khan et al. [2017]). Prostate cancer is the formation of malignant tumors where the mutation of prostate gland's normal cells take place (American Cancer Society [2019]). According to Altaklinik, factors contributing to prostate diseases may be related to aging, family history, lifestyle, effects from other diseases, etc. While each of these prostate diseases would cause complications to the male patient, non-other than the prostate cancer that would leave a life-threatening impression. It is of utmost clinical significance for domain experts to detect prostate cancer at an early stage to increase the survivability rate of a patient. In the survey by Smith-Palmer et al. [2019], it is pointed out that prostate cancer has been one of the most reported cancer disease among the male population in the world associating with high mortality rate as well as health burden to the respective patients. According to the cancer incidence and mortality patterns in Europe 2012 by Ferlay et al. [2013], prostate cancer accounts for 23% (417,000 cases) of cancer associated with male patients; of these, 10% (92,000 cases) result to cancer-related fatalities. According to Smith-Palmer et al. [2019], it is found that deaths among men directly caused by prostate cancer are fewer compared to deaths due to complications from prostate cancer. Bell et al. [2015] reported that the incidental prostate cancer prevalence at the time of death range from 5% for age <30 to 59% for age >79.

Because of these staggering numbers of prostate cancer occurrence among the male population, early detection of prostate cancer is not trivial. Miller et al. [2003] stated that two-third of the diagnosed prostate cancer patients do not display related symptoms. Generally, there are two screening phase examination for the diagnosis of an prostate cancer (Naji et al. [2018]). The first phase involves the screening of the elevation of prostate-specific antigen (PSA), which is an enzyme produced specifically by prostate cells. While increment of PSA level is a vital biomarker that indicates potential prostate cancer threat, at the same time, it indicates the presence of other prostate diseases (American Cancer Society [2019]). Prior to the elevation of PSA levels, the male patient would be subjected to digital rectal examination (DRE) or biopsy for the confirmation of cancer. While PSA does not serve as a strong indicator for the presence of prostate cancer, DRE has been reported to be ineffective by Naji et al. [2018]. On the other hand, biopsies is an invasive method that refer to the direct sampling of a prostate tissue using a special needle by pathologist (Patel and Jones [2009]). Investigation by Smith-Palmer et al. [2019] stated that such detection and diagnosis of prostate cancer has raised concerns regarding over-treatment and add on unnecessary burden upon the already overstretched healthcare systems. Not to mention that such treatment would increase the potential for side-effects such as impaired bowel movement, impotence, frustration of patient partner, etc. which would impact the quality of life at a patient level (Smith-Palmer et al. [2019]). There is a dire need to push boundaries for better diagnosis and detection given a prostate cancer.

A study by Bonekamp et al. [2011] demonstrates the advancement of multi-parametric magnetic resonance imaging (MRI) for better detection in identifying prostate cancer. Domain experts in Altaklinik verify multi-parametric MRI as the most reliable imaging procedure for early detection of prostate cancer. There are multiple critical information that could be retrieved from multi-parametric MRI to support the decisions of an urologist, which includes, detection of a malignant tumor, spatial information of a tumor, size and volume of a tumor, etc. (Altaklinik). Concurrently, several attempts have been made by researchers to develop a hybrid (human in the loop) and fully automated systems that would support the decision making of human in diagnosing prostate cancer (Wang et al. [2017], Liu et al. [2017a], Reda et al. [2018], Lucas et al. [2019], Nagpal et al. [2019]). Such system requires various techniques and implementations for the development, where early success and breakthrough results could be observed from machine learning algorithm in computer vision (Tătaru et al. [2021]).

However, with the advancement of computational resources and the exponential growth of data availability, a novel approach, called deep learning, emerges to address the shortcomings of machine learning whilst providing state of the art performances and enhancements. Wang et al. [2017] stated that deep learning methods aims to address the shortcomings of machine learning methods that depend on handcrafted feature extraction, suitable only for specific medical-image analysis use cases. Despite the capability of a deep learning model to learn feature extractions and optimise the cost function, conventional diagnostic model trained on softmax classifier (Wang et al. [2017], Liu et al. [2017a]) only learns a mapping from images to labels, such that different labels are adequately distant enough in the feature space to form a

decision boundary. Deep metric learning attempts to learn an embedding from deep neural networks such the data representation would learn similarities from the data set via distance metric.

Deep metric learning is able to effectively enhance discriminative power between class clusters for better classification capability. Chopra et al. [2005] proposed contrastive loss, one of deep metric learning techniques with the aim to imposed stricter constraints between different labels such that various labels are discriminative enough in the feature space, viz. images with the same labels and contents that are close to each other, and vice versa. In recent years, another popular deep metric loss, triplet loss, was proposed by Schroff et al. [2015] to overcome the shortcomings of contrastive loss. Contrastive loss and triplet loss however suffer from sampling strategy where pairwise samples or triplets generation are crucial for the model to learn. Furthermore, contrastive loss and triplet loss are not able to explicitly perform classification task without relying on pseudometric, such as k-NN (Zhou et al. [2020], Sundgaard et al. [2021], Pal et al. [2021]). In order to map a feature embedding to class labels, deep metric metric methods with softmax classifier, which is an extension of softmax loss was introduced. The drawback of the original softmax loss is that it does not possess margin criterion for optimisation of the feature embedding to enforced intra-class similarity and inter-class discrepancy. This inspired a series of softmax losses that incorporated margin to enhance intra-class compactness and inter-class diversity, known as margin-based softmax loss. Most notably, margin-based softmax loss (Liu et al. [2017c], Wang et al. [2018], Deng et al. [2018]) has achieved state-of-the-art performance in face recognition and face verification task. Liu et al. [2017c] stated that learning Euclidean distance in the cartesian space of an embedding from the original softmax loss is not optimum and proposed angular softmax loss function, such that instead of Euclidean distance the geodesic distance could be learnt in a sphere space for better feature representation. Angular softmax loss adapted multiplicative margin which suffers from the inconsistency of margin boundaries for different classes. This approach however, inspired a series of state-of-the-art margin-based softmax loss that implemented additive margin for better performance with outstanding breakthroughs in the field of deep metric learning, most notably, CosFace loss by Wang et al. [2018] and ArcFace loss by Deng et al. [2018].

On the other hand, whilst the rich information retrieved from MRI images may enable an urologist to detect cancerous lesions as well as evaluate its properties, such repetitive task would be laborious in a long-run and prone to human-errors leading to unimaginable consequences. This is especially challenging and time consuming for urologist to distinguish uncertainties during specimen viewing as well as to seek guidance from confirmed similar images or cases. Content based image retrieval (CBIR), was introduced with the intention to use images as an input to search for similar use cases (Dubey [2021]). Ni et al. [2017] stated that similarity between patient pairs with key clinical representations could be derived via the distance metric of deep metric learning in medical application. Subsequently, margin-based softmax loss could enhance the capability of CBIR through embedding optimisation.

1.2 Research Objectives

While it is true that mpMRI imaging would aid domain experts to detect cancers and reduce mortality rates, there exists the issue of false positive biopsies that would lead to over-diagnosis. In practice, when a lesion is deemed to be suspicious from the imaging, biopsy is often associated. This thesis hypothesises that through the use cases of deep metric learning, specifically margin-based softmax loss, the model is able to learn an embedding distance by enforcing similarity for intra-class samples and dissimilarity for inter-class samples in prostate disease classification among patients. Patient similarity classification is able to draw a discriminate decision boundary for the detection of clinically significant prostate cancer. In turn, this discovery would act as a critical clinical decision support for biopsies and prevent over-diagnosis. In addition to prostate cancer classification, the sub-optimal task of this thesis is to exploit the capability of CBIR on volumetric medical images with margin-based softmax loss. Through the implementation of CBIR, the thesis aims to investigate the extension of margin-based softmax loss on the context of visual similarity in mpMRI, such that the method could gain insights related to the distance metric as a measurement of patient pairs represented by learned features.

1.3 Thesis Structure

The structure of the thesis is as follows. In chapter 2, a summary about related works and literature review will be provided. Chapter 3 attempts to elaborate the building blocks of related knowledge and techniques used in this thesis. Chapter 4 explains the set up of ablation studies as well as the mathematical concepts that support this thesis. Chapter 5 presents the discussions and analyses of the ablation study results. Lastly, Chapter 6 would provide summary of the thesis works and potential future work for improvements.

2. Related Work

This section provides an overview of literature reviews and relevant works that have been conducted by researchers in the past and recent that could contribute to meaningful insights for the thesis.

2.1 Deep Learning for Medical Applications

Deep learning is a subset of machine learning that renders a deep neural network to train in an end to end fashion through backpropagation algorithm. Emergence of deep learning has taken over the artificial intelligence community by storm by leveraging the need of feature engineering as opposed to machine learning techniques (Wang et al. [2017]). Convolutional neural network (CNN) was proposed by LeCun et al. [1989] which further pushes the boundary of neural network to learn a gird-like topology data set, and ever since, CNN has beeen the *De Facto Standard* for computer vision applications. CNN layers spark an avalanche of very deep neural network architectures, most notably, VGG16 (Simonyan and Zisserman [2015]), GoogleNet (Szegedy et al. [2014]), ResNet (He et al. [2015]), DenseNet (Huang et al. [2016]) and many more that achieved outstanding performance in the ImageNet challenge (Deng et al. [2009]).

Although ImageNet database consisted of image objects from natural scenes in the real world, the implementations of deep learning techniques are adapted by the medical community. In medical applications, deep learning has been widely adapted in image segmentation tasks and disease classification tasks. UNet was introduced by Ronneberger et al. [2015] for image segmentation task where multiscale features are extracted through the downsampling process followed by skip connections connecting the features at each upsampling process. Such U-shape model structure has gained traction in inspiring similar model architecture (Chen et al. [2016], Oktay et al. [2018], Lee et al. [2020]) that are predominant in the brain tumour image segmentation (BraTS) challenge (Menze et al. [2015]). Liu et al. [2017b] proposed an-isotropic ResNet to tackle the issue of low depth to width or height ratio where conventional networks assume an isotropic(cubic-like) volumetric image. Cheng et al.

[2016] adapted stacked denoising autoencoder (SDAE) network architecture to detect benign and malignant breast lesions from computed tomography(CT) scans. During the Covid-19 pandemic, deep learning techniques has been presented by researchers (Zhao et al. [2021], Zhong et al. [2021], Sriram et al. [2021]) to identify Covid-19 patients through pneumonia detection via chest radiograph(CXR) images.

2.2 Computer-aided Diagnosis for Prostate Diseases

Prostate is an important reproductive organ for male and prostate diseases could bring about inconvenience to the patients' daily life. Particularly, prostate cancer has been one of the most reported cancer diseases in the male population with high risk prostate cancer possessing life-threatening potential (Chang et al. [2014]). Computer aided diagnosis attempts to build an automated system with deep learning techniques with the aim to support clinical decisions for prostate diseases. Wang et al. [2017] demonstrated the capability of deep learning in contrast to machine learning algorithm for fully automated prostate cancer 2D-MRI imaging classification. Concurrently, Liu et al. [2017a] introduced XmasNet, a VGG net inspired model, for prostate cancer lesion classification with multiparametric MRI (mpMRI). In the work by Reda et al. [2018], late fusion of features from prostate specific agent (PSA) and DWI MRI images are fused as an input into deep neural network for benign and malignant prostate tumour classification. The work by Reda et al. [2018] is however not end-to-end and relies on preprocessing techniques, such as KNN classifier and non-negative matrix factorization for feature extractions. Lucas et al. [2019] proposed automatic Gleason pattern classification using deep learning for Gleason grade group (GGG) determination of prostate biopsies. Lucas et al. [2019] employed inception network on prostate histopathology image for GGG classification. On the other hand, Nagpal et al. [2019] developed a two-stage deep learning system for Gleason scoring of prostate cancer using prostate histopathology image. Stage one involved segmentation task of Gleason pattern derive from prostate histopathology image. Whereas, stage two employed classification of Gleason scores with the features learned from the model in stage one. Pellicer-Valero et al. [2021] presented an automated framework that provides detection and lesion segmentation of prostate cancer with mpMRI images as well as Gleason score estimation on ProstateX and Valencia Oncology Institute Foundation (IVO) dataset.

2.3 Deep Metric Learning

Deep metric learning is a technique that implements distance metrics on deep neural network to enforce intra-class compactness and promotes inter-class repulsion. Conventional classification model utilise softmax loss, i.e. softmax output with cross entropy loss that attempts to learn a mapping from data to labels through feature separability without optimising feature similarity. Deep metric learning attempts to leverage the shortcomings of softmax loss by learning an optimised embedding space through distance metrics such that objects with the same labels are close to each other, whereas objects with different labels are far apart. Deep metric learning has been successfully applied in face recognition (Schroff et al. [2015], Liu et al. [2017c], Wang et al. [2018], Deng et al. [2018]), person re-identification (Hermans

et al. [2017], Chen et al. [2017]), information retrieval (Zhong et al. [2021], Ni et al. [2017], Onga et al. [2019]), etc. Inspired by spring model for repulsion and attraction, Chopra et al. [2005] introduced contrastive loss by optimising pairwise objectives on euclidean distance. In contrastive loss, objects with same labels should have minimal pairwise distance, whereas objects with different labels should have maximal pairwise distance given a margin. Recently, Schroff et al. [2015] introduced triplet loss such that, triplets consist of an anchor, positive and negative. Positives are matching thumbnails with the anchor, in contrast to the negatives, which have different labels as the anchor. Triplet loss aims to minimise distance margin of the positive pairs and maximise the distance margin of the negative pairs. Selections of triplets are vital for smooth learning of triplet loss. Hermans et al. [2017] investigated the criterion of triplets sampling and found out that batch hard approach, where hard positives (positives furthest from the anchor) and hard negatives (negatives closes to the anchor) yield the best results. Kumar et al. [2017] proposed smart mining strategy for triplet loss, but it is only possible to do the triplet sampling process offline without on the fly, which is computationally expensive and not efficient. Extending the work of triplet loss, Chen et al. [2017] introduced quadruplet loss by attempting to make intra-class variation smaller, while inter-class variation larger, as opposed to triplet loss that neglect the features class variation. Contrastive loss, triplet loss and quadruplet loss suffer from the need of good sampling strategy to prevent class collapse (Levi et al. [2020]), where the model learn that all features should be the same label for a minimum loss. Magnet loss was proposed by Rippel et al. [2015] to operate on the data set distributions as a whole instead of dealing with pairs, triplets or quadruplets. However, magnet loss did not gain much attention due to the difficulty of scaling and complexities.

While contrative losses (contrastive, triplets, quadruplets and magnet loss) suffer from sampling strategies and complexity issue, taking a step back, softmax loss could be qualified as a deep metric learning (Chan [2021]). The main concern about softmax loss is the ability to enact discriminative features between dissimilarities. Wen et al. [2016] attempt to address this shortcoming through center loss, which acts as a regularisation term for softmax loss to enforce class clustering around the defined class center. Center loss however suffers for the need to provide the number of class centers for ideal learning. To address the lack of discriminative power in the softmax loss, the lack of margin criterion drove researchers to research on margin-based softmax loss for better embedding distance learning. Liu et al. [2017c] discovered that optimising an euclidean space might not be the best approach and proposed that optimising the geodesic distance in an hypersphere embedding could be beneficial, such that the class centers have the same radius to the center. In the work by Liu et al. [2017c], cosine distance between the features and the class center are computed by taking the product of normalised weights and normalised feature inputs, which are used as inputs to softmax loss. A multiplicative margin is implemented on the angle between the features and the class center, which formulates angular softmax loss. Angular softmax loss however suffers from inconsistency in the decision margin since the loss function depends on the cosine angle. Nonetheless, the formulation of angular softmax loss inspired a series of methods with angular distance as a metric learning, commonly known as margin-based softmax loss (Deng et al. [2018]). To address the drawbacks of angular softmax loss, additive margin penalty has been

introduced. Most notably, Cosface loss by Wang et al. [2018] with additive margin on the cosine distance and Arcface loss by Deng et al. [2018] with additive margin on the cosine angle. Both Cosface and Arcface bear a lot of resemblance, which the former method yield marginal superior results for face recognition task.

2.4 Deep Metric Learning for Medical Applications

Deep metric learning has proven its capability in face recognition tasks due to the implementation of distance metric to produce discriminative features for visual similarity. This core idea could be transferred to medical applications where deep metric learning attempts to learn clinically significant distances that support better decision making. In the work by Sundgaard et al. [2021], the authors proposed an automatic diagnostic algorithm for detecting otitis media using deep metric learning. Sundgaard et al. [2021] implemented contrastive loss, triplet loss and multi-class N-pair loss for deep metric learning, then incorporated k-NN on feature embedding for classification of otitis media. Similarly, in another work by Pal et al. [2021], the authors employed deep metric learning for cervical cancer classification. Pal et al. [2021] adapted batch-hard triplet loss, contrastive loss and N-Pair embedding loss with k-NN for cervical image classification.

Other than diseases diagnosis and classification with deep metric learning in medical applications, medical images analysis could be carried out on the basis of content based image retrieval (CBIR) system. The core idea of CBIR relies on image as a query to retrieve information from large database such that the need to perform query by keyword could be leveraged (Zin et al. [2018]). Ni et al. [2017] presented a work to investigate patient similarity with deep metric learning from electronic health records (EHRs). Ni et al. [2017] use quadruplet loss as a deep metric learning framework for fine-grained disease classification. Meanwhile, Onga et al. [2019] adapted CBIR with deep metric learning on brain MRI images. They implement 3D convolutional autoencoder to learn a latent space for feature extraction from 3D brain MRI images. This low dimensional representation is then used as feature matching with deep metric learning to retrieve similar cases corresponding to the brain MRI images. Recently, due to the resurgence of covid-19 cases and the availability of abundant chest radiograph (CXR) images, Zhong et al. [2021] attempt to use deep metric learning technique for CBIR on CXR images such that the retrieved image could provide meaningful clinical information instead of using keywords as queries. In the work by Zhong et al. [2021], they adapted multi-similarity loss by Wang et al. [2019], associating it with attention mechanism to classify the severity of pneumonia cases given CXR images.

3. Background

This chapter aims to explain the building blocks required for this thesis work. In this chapter, anatomy of the prostate as well as medical image modalities from multi-parametric MRI will be elaborated. Concepts and implementations of deep learning techniques are also introduced, with the advantages and disadvantages being discussed. Finally, deep metric learning methods will be contextualized for a better overview on how state of the art method would support the decision making for prostate cancer.

3.1 Prostate

Prostate is an important organ of a male reproductive systems, located at the center of the pelvis, below the bladder and in front of the rectum, as shown in Figure 3.1. Concurrently, the prostate also surrounds part of the urethra. Generally, the prostate weighs around 20 grams and resemble the size slightly larger than a walnut.

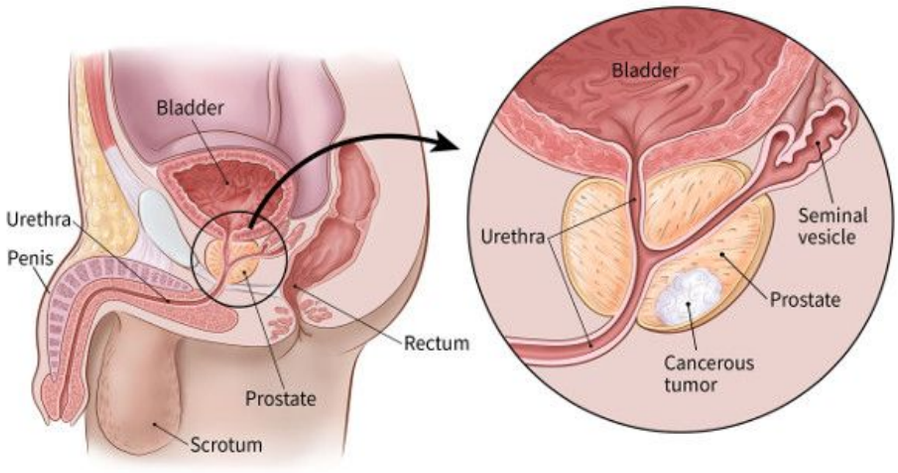


Figure 3.1: Location of prostate as well as it's surrounding organs and glands. Image source from American Cancer Society [2019].

3.1.1 Prostate Function

Prostate is a small gland consisting of several secretory glands, surrounded by collagen fibers and muscle. The main function of the prostate is to generate a viscous fluid that serves as nourishment and protection for the sperm cells, which make up part of the semen (InformedHealth.org [2011]). Semen also consists of seminal fluid supplied by the seminal vesicles, which is a pair of glands situated above the back of the prostate. During an ejaculation, the sperm cells produced by the testicles, together with prostatic and seminal fluid are transported to the urethra. Prostate is also able to generate specific enzymes, such as prostate specific antigen (PSA), which may serve as a key bio-marker for the indication of prostate related diseases through the monitoring of PSA levels elevation (Altaklinik).

3.1.2 Prostate Zones

There are three anatomical zones within the prostate (InformedHealth.org [2011]).

- **Transition zone**: The zone that surrounds a fraction of the urethra and by-pass the prostate. This is the zone where enlargement of glandular tissue is commonly occur.
- **Central zone**: The zone that wraps around the ejaculatory ducts and position behind the transition zone.
- **Peripheral zone**: The outermost zone of the prostate.

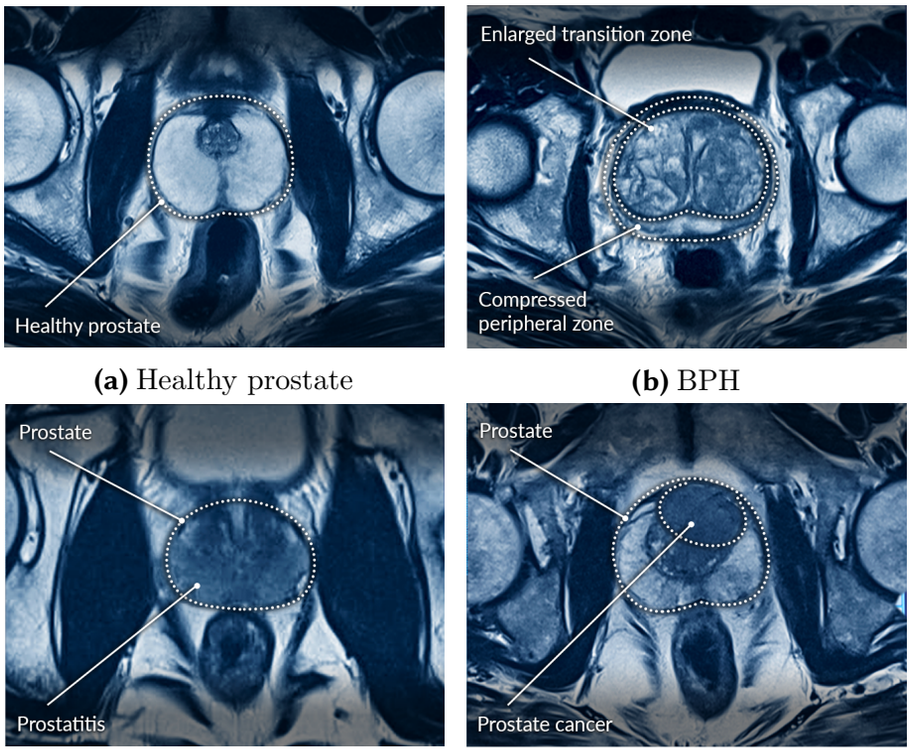
Progressive enlargement of a prostate zone would inevitably compress the remaining zones. As shown in Figure 3.1, due to the anatomical position of the prostate between the bladder and urethra, prostatic zone enlargement would typically lead to urinary dysfunction. Some common prostate diseases resulting from prostate zone enlargement would be discussed in section 3.1.3.

3.1.3 Prostate Diseases

In section 3.1.1, it is stated that elevated PSA levels could signify prostate diseases. Some of the most common prostate diseases include benign prostatic hyperplasia (BPH), prostatitis and prostate cancer. Figure 3.2 shows a schematic view of MRI images depicting the state of prostates in various prostate diseases as compared to a healthy prostate.

3.1.3.1 Benign Prostatic Hyperplasia (BPH)

BPH is a prostate disease that refers to the enlargement of prostate, commonly occur by men in the middle age or above (Roehrborn [2005]). Such enlargement, typically in the transition zone of the prostate would further compress the peripheral zone, causing male reproductive dysfunction. This phenomenon, referred as excess tissue growth, is due to the formation of benign nodules in the area of the prostate, caused by the proliferation of epithelial and stromal cells. As men age, the discrepancy of hormones contributed to the growth of excess tissue.



(a) Healthy prostate **(b)** BPH

(c) Prostatitis **(d)** Prostate cancer

Figure 3.2: MRI images of (a) healthy prostate, as well as prostate diseases: (b) BPH, (c) prostatitis and (d) prostate cancer. Image source from Altaklinik.

3.1.3.2 Prostatitis

Prostatitis is a prostate disease characterised by swelling as well as inflammation of a prostate gland (Khan et al. [2017]). At some point in life, men may or may not experience prostatitis. Mild symptom, such as chronic prostatitis, may come and go in a few months, without being potentially life-threatening. Acute prostatitis is more rare and potentially life-threatening, which is caused by bacteria infection in the premises of prostate gland.

3.1.3.3 Prostate Cancer

Prostate cancer occurs when normal cells in the prostate gland mutate, gradually form malignant tumours. According to American Cancer Society [2019], majority of the discovered prostate cancers belong to adenocarcinomas, where the cancers propagate from the gland cells, which are the cells that produce prostatic fluid mixed in the semen. Prostate cancer could be daunting and potentially life-threatening; however, not all discovered tumours are malignant. The tumour could turn out to be benign, where the tumour cells have a distinct borders and do not invade the surrounding tissue. In case of a malignant tumour, the degree of malignancy would determine the potential risk of the tumour and will be discussed in section 3.1.4. The primary purpose of this thesis work focus on detecting clinically significant prostate cancer via MRI imaging of a patient, such that the supporting decision of invasive prostate cancer could be defined.

3.1.4 Prostate Cancer Diagnosis

3.1.4.1 Prostate Specific Antigen (PSA)

Detection of early stage prostate cancer is significantly challenging as the prostate tumour cells do not spread significantly at an early stage, where the identification of tissue enlargement is not justified. In the aforementioned section 3.1.1, fortunately, the monitoring of PSA level in the patient blood, serves as an early warning for potential prostate diseases, in case of elevated PSA level. According to Altaklinik, elevated PSA level is defined as a patient who has PSA level that raises continuously or spike within a period of 6 months, with a PSA level threshold of >4ng/ml.

3.1.4.2 Multi-parametric MRI (mpMRI)

It has been discussed in section 3.1.3 that elevation of PSA levels would not specifically pinpoint the type of prostate diseases. Therefore, patients with elevated PSA levels are advised to undergo MRI scan to detect anomalous tissue. Multiparametric prostate MRI enable the doctor to identify the diseases of the prostate as well as the condition of the disease via the analysis of generated image. More importantly, prostate MRI scan is particularly effective for the initial diagnosis of prostate cancer, where MRI images enables doctor to retrieve the positional attributes as well as the spatial properties of prostate tumours. Prior to mpMRI, Demirel and Davis [2018] stated that single sequence type MRI (T1W) was not able to the generate high tissue resolution images, which cause the discrepency between prostate and it's surrounding tissues. Next, mpMRI was introduced to leverage the shortcomings of single sequence MRI. In the context of Demirel and Davis [2018], mpMRI is essentially defined as the method to retrieve a genuine 3D prostate image through the combination of MRI image sequences. Typically, this combination consisted of image sequences, such as T2-weighted (T2-WI), diffusion weighted (DWI), dynamic contrast enhanced (DCEI), etc. In this work, Altaklinik provided MRI sequences of T2-W, DWI and apparent diffusion coefficient (ADC) as depicted in Figure 3.3.

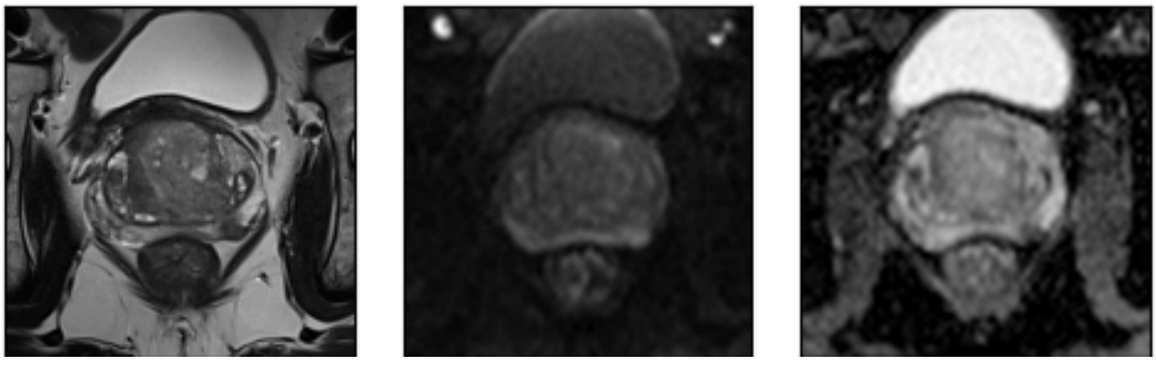


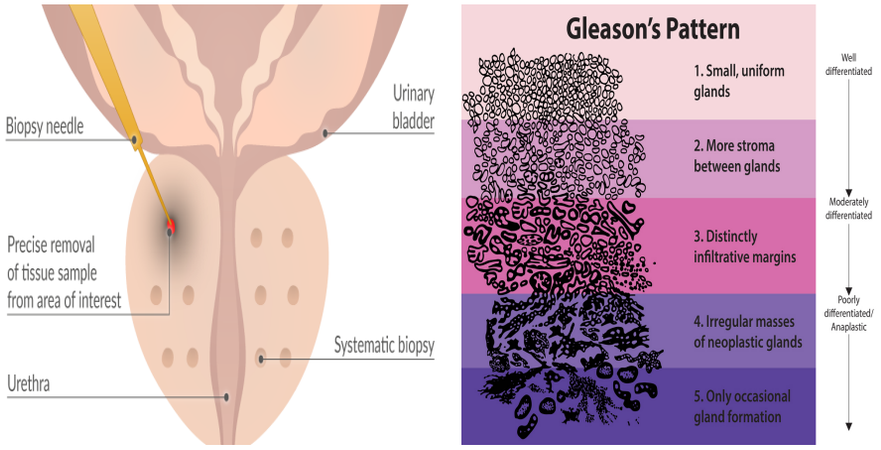
Figure 3.3: MRI sequences starting with (left) T2W, (middle) DWI, (right) ADC.

T2-W sequence is a MRI image that is derived through the calculation of water density in the tissue. T2-W is able to capture low signal intensity around the tumour tissues. According to several studies by Wu et al. [2012] and Yoo et al. [2015], T2-W sequences are not sufficient to identify cancerous tissues in transitional zone and central zone. On the other hand, diffusion-weighted sequences are affiliated with

the proton diffusion of water atoms which indicates the Brownian motion of water molecules in the tissues. Despite the low resolution of MRI images as compared to T2-W sequences, Lee et al. [2018] pointed out that DWI image sequences are able to display better central zone and transitional zone tumours. ADC image sequences or ADC maps belong to the diffusion-weighted sequences. In ADC image sequences, Demirel and Davis [2018] stated that clinically significant prostate cancers appear hypointense, as compared to non-cancerous tissues.

3.1.4.3 Biopsy

In case of potential detection of prostate cancer from MRI images, biopsy is necessary to confirm the presence of prostate cancer via sampling of tissues around the region of interest and examine by pathologist. Figure 3.4 (a) depicts a schematic figure of how biopsy is carried out.



(a) Biopsy Process **(b)** Gleason Pattern

Figure 3.4: (a) Sampling of cancerous prostate tissue with biopsy needle. Image source from Altaklinik. (b) Gleason pattern from the pathological image of prostate tissues. Image source from National Cancer Institute.

Once biopsy is carried out, pathologist would need to assess the degree of prostate tumour malignancy. The degree of malignancy would be attested by pathologist according to the Gleason's pattern as depicted in Figure 3.4 (b) and graded according to the Gleason grading system as shown in Table 3.1. The higher the Gleason score, the higher the malignancy of a tumour in the prostate. For malignant tumour, the lowest Gleason score is 6 and the highest Gleason score is 10. A new Gleason grading schema, namely, Gleason grade group (GGG), was introduced by the International Society of Urological Pathology (ISUP) (Epstein et al. [2016]), such that the lowest grade is 1 instead of 6 in the formal Gleason score schema.

Deciphering from the context of Table 3.1, 50% of male patients being diagnosed with prostate cancer possess tumour of Gleason group 1, qualified as low risk. Whereas, 25% of prostate cancer patients are likely to experience tumours with Gleason group 2 and 3, qualified as medium risk. Whilst the rest of the 25% prostate cancer patients accountable for Gleason group 4 and 5.

Gleason Grade Group (GGG)	Gleason Score (GS)	Prognosis	Frequency
1	6=3+3	Low Risk	50%
2	7a=3+4	Medium Risk	25%
3	7b=4+3		
4	8=3+5	High Risk	25%
	8=5+3		
	8=4+4		
5	9=4+5		
	9=5+4		
	10=5+5		

Table 3.1: Gleason grading system displaying the classifications of respective Gleason grade group (GGG) or Gleason score (GS) corresponding to the Gleason's pattern, prognosis and frequency.

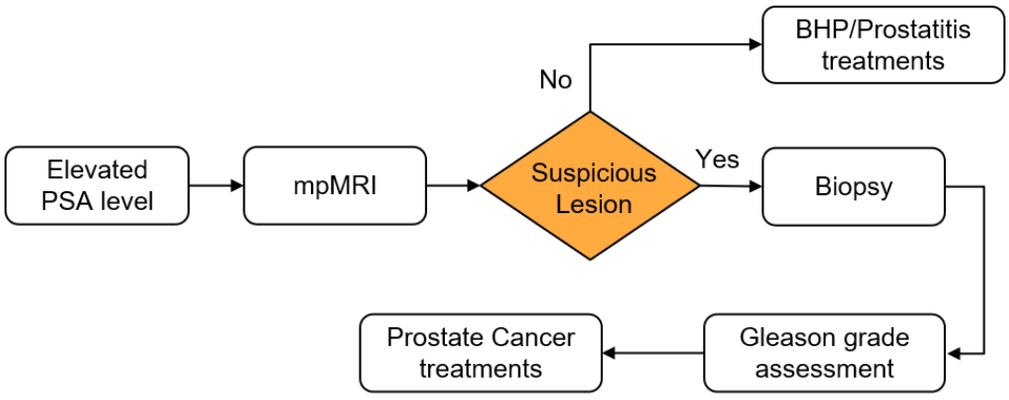


Figure 3.5: Schematic flowchat about the pathway from prostate disease diagnoses to corresponding treatment.

3.1.4.4 Treatment Pathway

Figure 3.5 illustrate a flowchart summarising the pathway of prostate diseases detection to it's treatment. Elevated PSA level of a patient would indicate potential occurrence of prostate diseases. mpMRI may then be done for detail inspections to rank out the possibilities of prostate cancer. If suspicious lesion is detected, the patient would need to undergo biopsy to determine the Gleason grade of the lesion. Post Gleason grade results would support the decisions of the doctor to provide corresponding prostate cancer treatments. Else, the patient would be subjected to BHP or prostatitis treatments.

3.2 Deep Learning

Artificial Intelligence (AI) is defined as synthetic intelligence generated by machine that mimics or resembles biological intelligence, those displayed by humans or animals. According to Honavar [2014], the primary goal of AI research aims to enhance human understanding of reasoning, learning, creative processes, perceptual and linguistic through the gathering of knowledge. Therefore, in the early stage of AI, the system could be used to as a problem solver for tasks that are intellectually difficult for

humans, as mentioned in Goodfellow et al. [2016]. These tasks are, however, less complex for computers with mathematical constraints. Hard-coding knowledge for an AI system is not a trivial task which led to the need for the system to acquire their own knowledge. This process of knowledge acquisition could be achieved through pattern extraction from raw data, whilst the algorithm is generally known as machine learning.

Machine learning methods enable computer to learn from experience through hierarchical concepts such that each concepts is related to simple concepts. The approach of learning from experience leverages the need for human to manually handcraft knowledge for the computer needs. Machine learning methods such as naive Bayes, logistic regression, decision tree, support vector machine (SVM), etc. have proven performance boost in a wide range of field, including computer vision, natural language processing (NLP), speech recognition, etc. However, machine learning algorithms heavily depend on the representation of the data, known as feature. Such dependence on representations would require feature engineering for different use cases, where it is challenging to understand what kind of feature extractions would benefit the system.

Deep learning leverage the need for handcrafted feature extractions by introducing representations that are built upon other simple representations (Goodfellow et al. [2016]). In short, deep learning solves the problem of representation learning through deep composite representation that is able to extract high-level or abstract features from raw data as opposed to conventional machine learning techniques. Deep learning enables learning of sophisticated concepts by building blocks of simpler concepts through nested hierarchy of concepts, as depicted in Figure 3.6.

3.2.1 Artificial Neural Network

As illustrated in Figure 3.7, machine learning is an approach in AI, whereas deep learning is a subset of machine learning, an implementation that allows computer to learn from experience and data. The building blocks of a deep learning model are known as feedforward deep network, multilayer perceptron (MLP) or artificial neural network (ANN). In this work, the terminology for the quintessential of a deep learning model is unified as artificial neural network (ANN). Explained in Goodfellow et al. [2016], ANN are named networks because they are formed by the compositions of mathematical functions that map some set of input values to output values. The computation of ANN was inspired by biological neural system, where the information propagate from one node to the other as shown in Figure 3.8(a). In Figure 3.8 the network is feedforward, where the information flows from inputs x through the functions in the hidden layers and to the predicted output, $\hat{y}$. $h_n^{(l)}$ denotes a hidden unit, where l is the number of hidden layers, n is the number of hidden units in a hidden layer. The number of hidden layers represents the depth of the model, whereas the number of hidden units in a hidden layer defines the width of the model. As illustrated, all the hidden units from the previous layer, $h_n^{(l-1)}$ are connected to all the hidden units in the next layer, $h_n^{(l-1)}$. The relations between the connections of hidden units are known as fully-connected and the layer is known as dense layer.

$$f(\mathbf{x}; \mathbf{w}, b) = \mathbf{x}^T \mathbf{w} + b \tag{3.1}$$

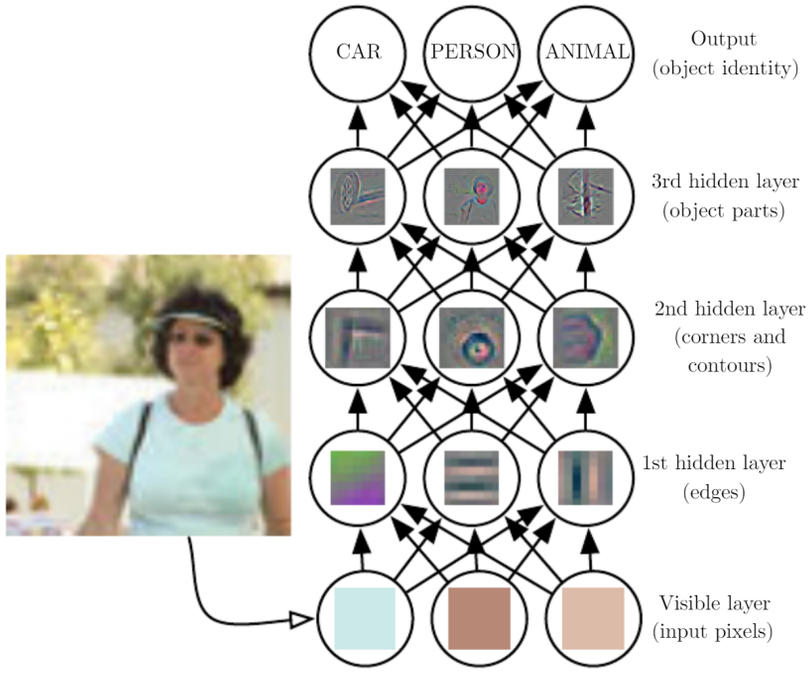


Figure 3.6: Visualization of deep learning representations through building complex concepts out of simpler concepts. On the bottom of the deep learning model, visible layer are raw inputs, which are the variables that the users could observe. The hidden layers consisted of variables that are not able to observe. The formation of the feature extractions became more complicated as the model learned in the deeper layers. Image source from page 6, Goodfellow et al. [2016].

$$\hat{y} = f(\mathbf{x}) = f^{(4)}(f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))) \tag{3.2}$$

Figure 3.8 (b) illustrates a direct acyclic graph that portrays a compact representation of an ANN. f is a function that maps an input value, where the most basic transformation is linear transformation, shown in equation 3.1, where $\mathbf{w}$ is a weight matrix and b is the bias term. The graph depicts the functions, f connecting together in a chain forming composite function, written as equation 3.2.

3.2.1.1 Activation Function

In section 3.2.1 the function f utilized a linear transformation in between layers. However, linear transformation of another linear transformation will only retain linear features and the model would only learn a linear decision boundary for a given task. Particularly when we need probability distribution as an output, linear mapping would not satisfy the condition. Therefore there is a need for non-linear transformation, which in the context of deep learning is called activation functions. Non-linearity properties derived from the activation functions would enable the network to learn non-linear decision boundaries through non-linear combinations of the weights and inputs.

Figure 3.9 depicts schematic plots about the most commonly used activation functions in the deep learning community. Rectified Linear Units (ReLU) implement the activation function shown in equation 3.3. Unlike linear unit, ReLU outputs zero across

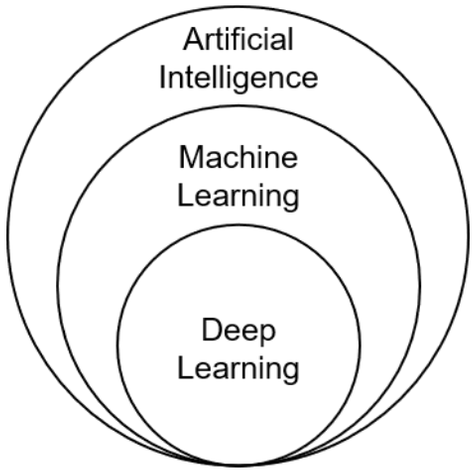
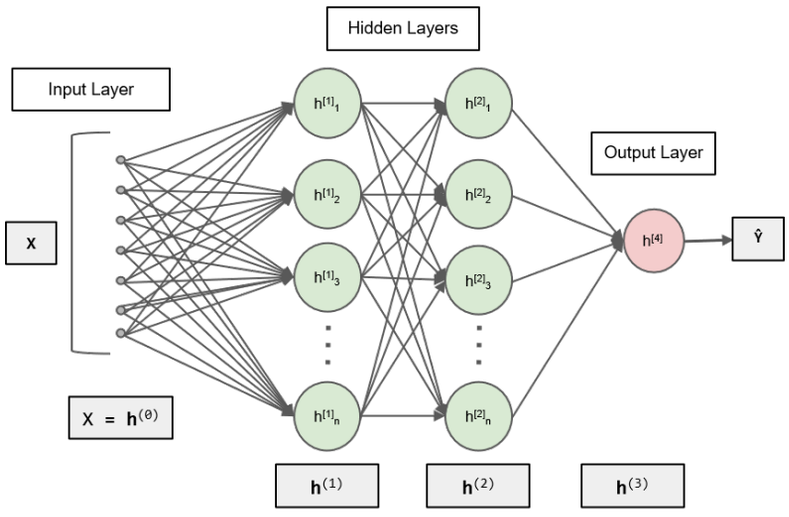
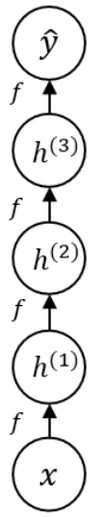


Figure 3.7: Venn diagram illustration about deep learning as subset of machine learning under the approach of AI.



(a) Artificial Neural Network(ANN)



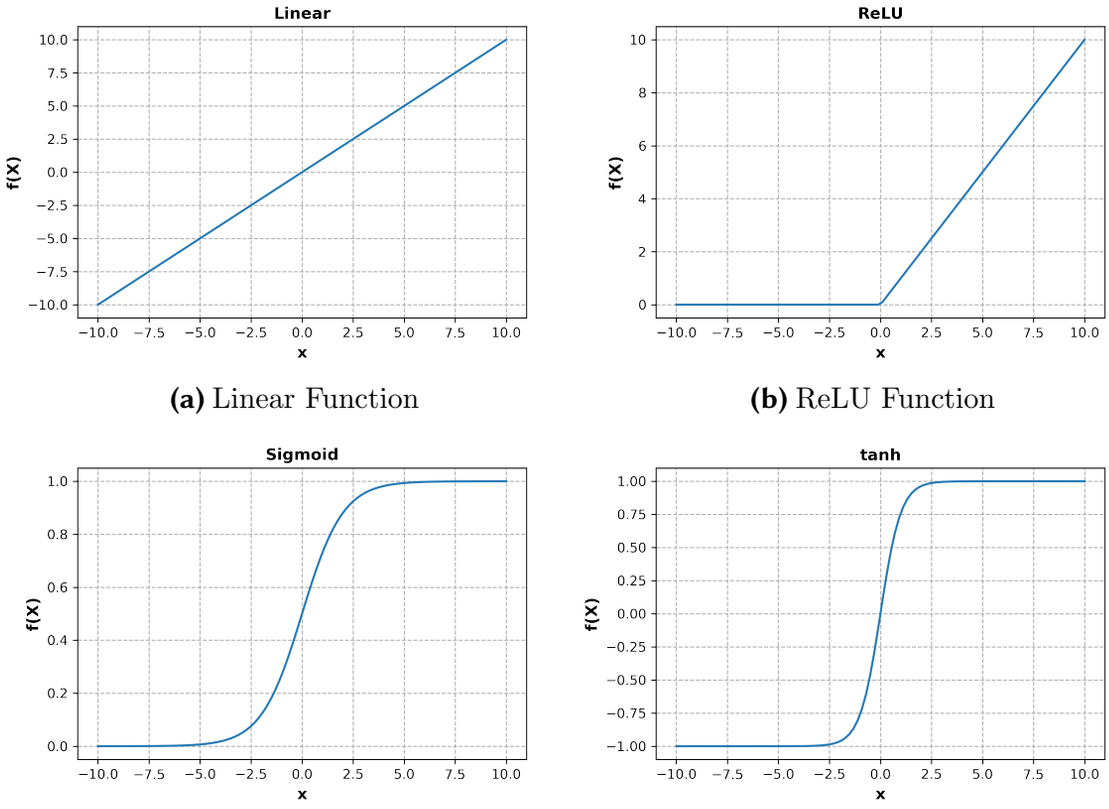
(b) Computational Graph

Figure 3.8: Illustration for an (a) ANN with 7 inputs, 2 hidden layers as well as 1 output. (b) depicts a computational graph as a representation of (a).

half its domain as shown in Figure 3.9 (b). ReLU also prevents saturation gradients which allows derivatives to pass through while training the model. Goodfellow et al. [2016] points out the drawback of ReLU is the gradient cannot be derived when their activation is zero, known as dead kernel.

$$f(z) = max\{0, z\} \tag{3.3}$$

Logistic functions such as sigmoid activation function, written in equation 3.4 and hyperbolic tangent activation function, written in equation 3.5, are closely related. As opposed to linear units and ReLU, logistic functions are saturated across most of their domain as illustrated in Figure 3.9 (c) and (d). When input value z is a large positive value, the function saturates to a high value and, on the flip-side, the function saturates to a low value when z is a large negative value. This property



(a) Linear Function

(b) ReLU Function

(c) Sigmoid Function

(d) tanh Function

Figure 3.9: Schematic plot of the most common used activation functions.

make logistic functions to be less fit as an activation function in between hidden layers. Whilst the use case of logistic functions fits the situation where the output requires a mapping in between [0,1] in case of sigmoid function or in the range of [-1,1] for hyperbolic tangent function.

$$\begin{aligned} f(z) &= \sigma(z) \\ \sigma(z) &= \frac{1}{1+e^{-z}} \end{aligned} \tag{3.4}$$

$$\begin{aligned} f(z) &= tanh(z) \\ tanh(z) &= 2\sigma(2z) - 1 \end{aligned} \tag{3.5}$$

Another popular choice of activation function commonly used for the output layer is softmax function. Softmax function maps the output of a classifier model, such that the output of the model resembles probability distrbution over n different classes. From equation 3.6, z_i is the unnormalized log probabilties of a model output. Shown in equation 3.6 the softmax function first takes the exponential operation of z_i and normalize the log probabilities to obtain the predicted label, $\hat{y}$. The outputs of softmax always sum to 1 such that the increment of the value in the unit would yield the decrement in the value for the rest of the unit, which led to a behaviour called winner-take-all (Goodfellow et al. [2016]).

$$z_i = logp(y = i|\mathbf{x})$$
$$softmax(\mathbf{z}_i) = \frac{exp(z_i)}{\sum_j^n exp(z_j)} \tag{3.6}$$

3.2.1.2 Cost Function

A cost function or loss function calculates the distance of the predicted output of the algorithm and the expected output which measures the performance of the model. The cost function of each model varies the objective tasks. In this work, the objective task is to draw a decision boundary of an given input which correlates to a classification task. Therefore, the cost function in this work is the negative log-likelihood, equivalently expressed as the cross-entropy between the predicted output and the training output distribution. The cost function is expressed in equation 3.7.

$$L(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N} log(p_{model}(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})) \tag{3.7}$$

3.2.1.3 Training a Neural Network

Goodfellow et al. [2016] stated that almost all deep learning models are based on this important algorithm, which is stochastic gradient descent (SGD). Deep learning algorithms often need large training sets for good generalization, but training on large data sets would be computationally expensive and subject to hardware constraints. To leverage this problem, SGD was built on the assumption computing gradient descent using small batches, such that we sample a minibatch from the training data sets for each step. The gradient, $\mathbf{g}$ is computed by taking the derivatives of equation 3.7, as shown in equation 3.8, where m is denoted as minibatch size.

$$\mathbf{g} = \nabla_\theta L(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{i=1}^{m} \nabla_\theta log(p_{model}(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})) \tag{3.8}$$

The parameter of the model, $\boldsymbol{\theta}$ is updated as shown in equation 3.9, where ϵ is the learning rate.

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon\mathbf{g} \tag{3.9}$$

The update rules for a deep learning model could be computed, through the chain rule of calculus, which is known as backpropagation. Take for example, the derivatives of the computation graph in Figure 3.8(b) which can be expressed in equation 3.10.

$$\frac{\partial \hat{y}}{\partial x} = \frac{\partial \hat{y}}{\partial h^{(3)}}\frac{\partial h^{(3)}}{\partial h^{(2)}}\frac{\partial h^{(2)}}{\partial h^{(1)}}\frac{\partial h^{(1)}}{\partial x} \tag{3.10}$$

3.2.2 Convolutional Neural Network

As opposed to conventional machine learning techniques, ANN has made a paradigm shift in the field of AI and laid the foundations of deep neural networks. However, ANN only works well with one dimensional data sets, where the input data needs to be unravelled such that each data point correlates to an input unit to be fed into ANN models. This approach will not benefit data sets with grid-like topology, such as time series data, colored image, volumetric image, etc. Since ANN layers are fully connected, the number of trainable parameters in the network would not scale well given a sufficiently large input. Optimising a large number of parameters is not trivial and the dilemma persist when the network goes deeper. Inspired by how human vision perceive surroundings with receptive field in visual cortex (Fukushima [1988]), a special kind of neural network, namely, convolutional neural network was introduced. Convolutional neural networks (CNN) was proposed by LeCun et al. [1989] which adapted a mathematical operation known as convolution for matrix multiplication in ANN.

$$f(\theta) = (x * w)(\theta) \tag{3.11}$$

In mathematical context, convolution is an operation on two functions to achieve a function, expressed in equation 3.11. In CNN terminology, x is refer as the input, w is referred as the kernel. The output resulting from the convolution of x and w is simply a matrix multiplication product, defined as feature map. A schematic illustration of 2D convolutional operation is depicted in Figure 3.10, where stride is the step size movement of the kernel, similar to sliding window horizontally and vertically along the gird like input. Analogously, this concept could be easily extended to 3D use cases.

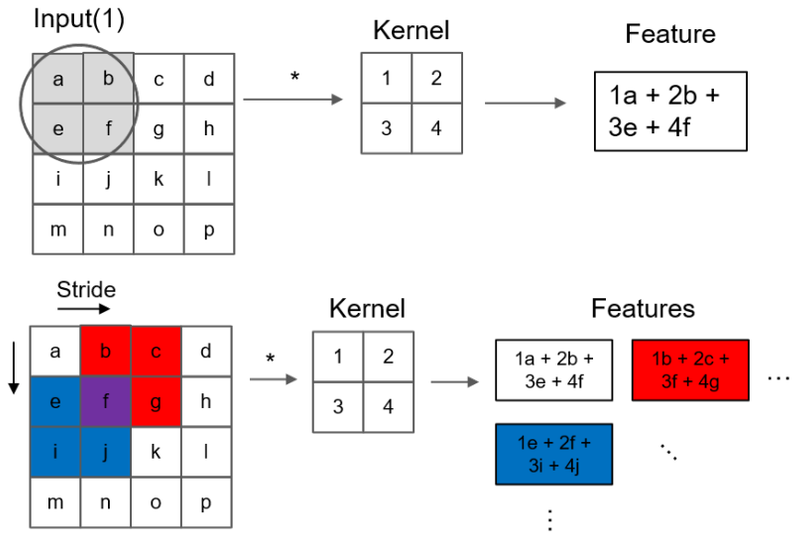


Figure 3.10: Illustration about 2D convolutional operation of a neural network. (Top) Input(1)∗Kernel a the matrix multiplication to obtain feature maps. (Bottom) shows the kernel moving along the inputs with a stride of 1 along the horizontal axis (red) and vertical axis (blue), "convoluting" with the region of interest.

Another essential feature of CNN is the capability to implement zero padding. The spatial representation of the feature, width and height, will shrink by one pixel less as compared to the input size at each layer. To prevent the diminishing of feature maps size, zero padding provides an efficient way to control input width and the size of the feature maps respectively (Goodfellow et al. [2016]). The output size of the feature maps after a convolution layer can be calculated with equation 3.12, where W is the width of the input, K is the kernel size, P denotes the number of padding and S is the number of strides.

$$Output = \frac{W - K + 2P}{S} + 1 \tag{3.12}$$

Generally, a pooling layer is implemented after a convolutional layer. A pooling function provides a kind of response summary along its nearby neighbor, typically with a kernel size of 2 and stride of 2. While a convolutional layer extends the feature maps along the depth, the width and height of the feature maps do not diminish drastically due to padding function. Pooling layer downsamples the feature maps along the width and height dimension, effectively reducing the number of parameter and only capturing important information at the region of interest. An example of max-pooling (Zhou and Chellappa [1988]) is depicted in Figure 3.11. An important property of pooling is that it enables learnt features to possess an approximate invariant to small translations given an input.

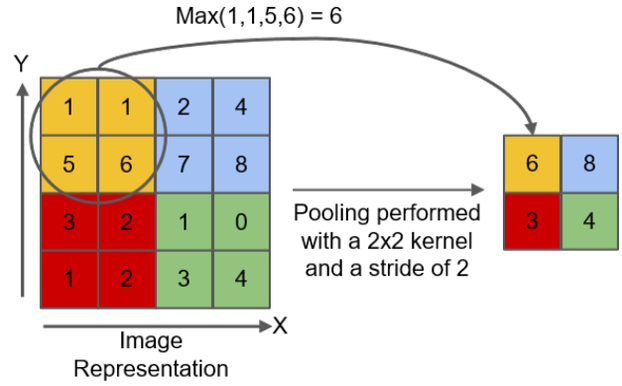


Figure 3.11: Max-pooling with kernel size of 2 and stride of 2 that takes only the maximum value of the region of interest.

Typically, a convolution block is made up of three layers. Firstly, a convolutional layer that performs several convolutions to achieve feature maps which is a set of linear activations. The next layer is the activation function layer that provides non-linearity transformation. The final layer is a pooling layer that reduce the width and height of the feature maps through summarizing neighbour responses.

Depicted in Figure 3.12, as opposed to fully connected dense layers, convolutional layers have sparse interactions due to their sparse connection resulting from the striding of kernel with well defined size. Subsequently, this property led to parameter sharing where the same parameter is distributed in more than one representation in the model. This property drastically decreases the number of training parameters in

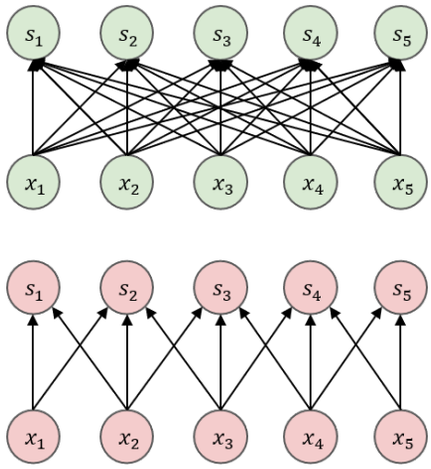


Figure 3.12: (Top) Densely connected network where all input **x** are connected to the units in the next layer **s**. (Bottom) Sparsely connected network where **s** is the product of convolution with kernel width of 3.

the machine learning system. Another important feature resulting from parameter sharing of convolutional layers is the manifestation of equivariance to translation. If the object in the input is shifted, the output would change in the same way.

3.2.3 Regularisation Techniques

The introduction of ANN and CNN has brought about paradigm shift in the field of AI to boost state-of-the-art performance in every aspect as compared to traditional machine learning techniques. Whilst this is the norm, the parameters in ANN and CNN models are substantially abundant and the model complexity are sufficiently complex, which contribute to the amount of degree of freedom for feature learning. Therefore, in order for ANN and CNN models to generalise well, a large amount of data are required to be fed to the model without concerns on overfitting, as shown in Figure 3.13. Underfitting refers to the model lacking the capability to manifest a good fit with the training samples. On the contrary, overfitting refers to the model lacking the capability to generalise well on the training model by learning the noise of the training data. A good fit is the most desirable scenario where the model is able to approximate the true function well.

To tackle the overfitting problem, several regularisation techniques where introduced.

Data augmentation is a technique for data generation by creating synthetic data on existing data via slight modifications. Data augmentation not only increases the data amount, it also aids the model to be robust against data noise, as such to help reduce overfitting while training deep learning models. Some common data augmentation techniques are rotation, random crop, translation, etc. (Shorten and Khoshgoftaar [2019]).

Dropout is a computationally inexpensive technique, proposed by Srivastava et al. [2014] to regulate the model. Since ANNS are based on multiple affine and nonlinear transformations, the idea of dropout could be demonstrated as removing a unit from the network by zeroing out the output value through zero multiplications as illustrated in Figure 3.14. This enables the model to learn a slightly different model at

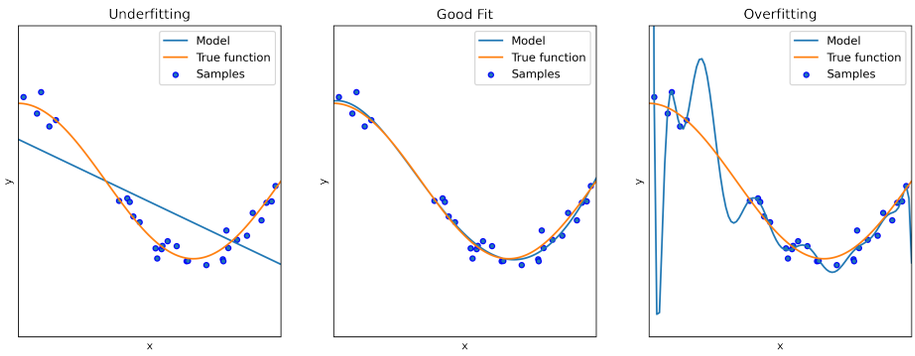


Figure 3.13: Linear regression plot with polynomial features of different degrees to approximate nonlinear functions. (Left) Polynomial of degree 1, (Middle) Polynomial of degree 4, (Right) Polynomial of degree 15. Plot adapted from Pedregosa et al. [2011] and reproduced for this work.

each iteration on the same data sets. While the dropout proposed by Srivastava et al. [2014] is known as element-wise dropout, there emerge several advanced dropout methods, such as spatial dropout Tompson et al. [2015], dropblock Ghiasi et al. [2018], weighted channel dropout Hou and Wang [2019], etc. These methods are beyond the the scope of this work and would serve as future references. This work uses the original dropout algorithm for all experiments.

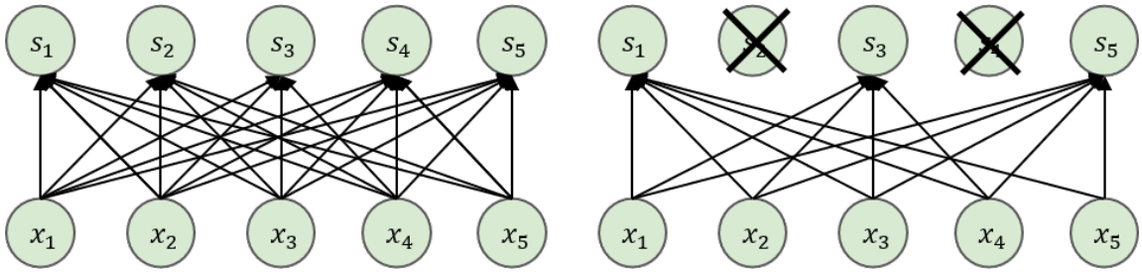


Figure 3.14: (Left) Fully connected network. (Right) s_2 and s_4 unit are remove through dropout algorithm where the connections from the previous layer are broken.

Early stopping is another inexpensive technique commonly adopted during the training of deep learning models. Early stopping means terminating the execution of the training step when the validation error is sufficient larger than the training error given a threshold throughout the training iterations. This prevents the model from severe overfitting and retains the model parameters with the lowest validation error.

Batch normalisation or commonly known as batch norm is a technique proposed by Ioffe and Szegedy [2015] to regulate and accelerate deep learning model during the training phase. The building blocks of batch norm are as illustrated in Figure 3.15 and summarised in the following five steps, Doshi [2021].

1. A_n are the mini batch activations from the previous layer acting as input for batch norm.
2. The mean and variance of the activations from the mini batch are computed.

3. The activations are normalised, such that the normalised values have unit variance and zero mean.
4. Normalised values are then scaled by γ and shifted with β. Where γ and β are both learnable parameters by the model, such that the model learns to optimise the best values.
5. Batch norm computes the exponential moving average of the mean and variance with momentum term, α. This implementation is especially important during an inference phase where moving average term is a good approximation for mean and variance of the training data.

The above steps implicitly provide an internal covariate shift in the batch norm layer for loss and gradient smoothing effect. This enables better convergence of the model computation, such that a higher learning rate could be adapted for a model to accelerate learning.

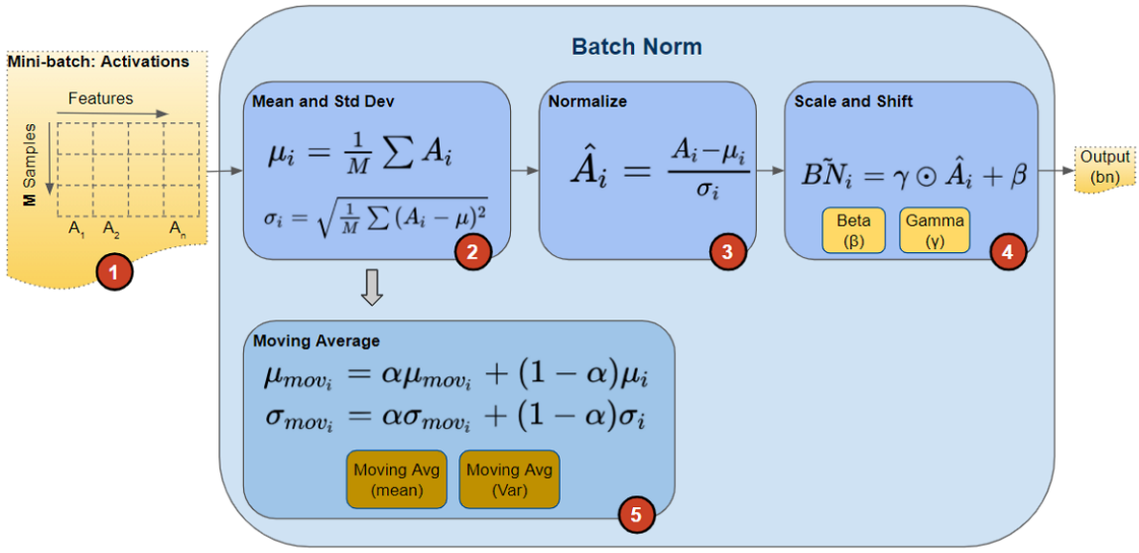


Figure 3.15: The building blocks of batch normalisation. The yellow boxes inside the batch norm algorithm are training parameters. Image source is a nice explanatory visual illustration by Doshi [2021].

3.2.4 ResNet

The introduction of CNN has led to breakthrough performance given an image classification task. This brought about an emergence of trend for deeper network architecture by stacking multiple layers to increase the model depth: for example VGG16 by Simonyan and Zisserman [2015], GoogleNet by Szegedy et al. [2014], ResNet by He et al. [2015], DenseNet by Huang et al. [2016], etc., have achieved stellar performance for their approaches in the ImageNet dataset challenge by Deng et al. [2009]. He et al. [2015] stated the degradation problem of deep networks where the accuracy of the models would be saturated as the model depth increase. He et al. [2015] also addressed the vanishing gradient problem of deep networks where signals from the deep layers diminish gradually as they reached early layers. In this work,

ResNet was chosen as the model due to its customizability from its model variants and outstanding performance without the trading off computational resources (Bello et al. [2021]). ResNet was able to tackle vanishing gradient of VGG16 and GoogleNet. DenseNet was inspired by ResNet with dense connections between layers resulting to increasing training parameters. In this work, we primarily focus on volumetric medical images analysis where 3D convolution layers would yield larger training parameters as compared to 2D operations. Hence, DenseNet is believed to be too computationally expensive for our considerations.

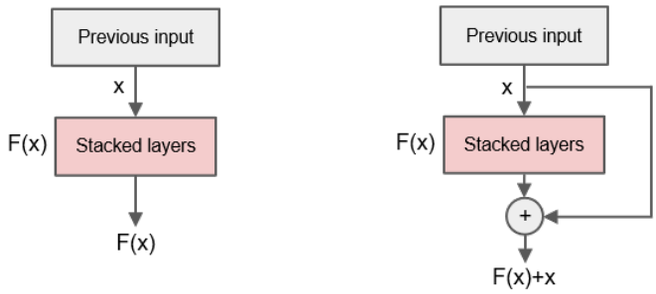


Figure 3.16: (Left) Network layers without shortcut connections. (Right) Residual learning of network layers with shortcut connections, where (+) is an addition operation. Image reproduced from paper by He et al. [2015].

The core idea of ResNet is the residual learning where there existed an identity mapping from the added layer from the shortcut connections or skip connections as shown in the right of Figure 3.16. Shortcut connections can skip one or more layers. Neural networks are good approximators such that by solving an identity function, the output of a function should be the input itself, as shown in equation 3.13.

$$F(x) = x \tag{3.13}$$

Intuitively, the network should be able to approximate an arbitrary function with additive input as shown in equation 3.14. Hypothetically, the network is able to easily push the residual, $F(x)$ to zero for an identity mapping.

$$F(x) + x = H(x) \tag{3.14}$$

Shortcut connections do not add extra parameter or computational complexity to the model, while backpropagation is able to bypass the shortcut connections, injecting signals to the shallow layers, effectively countering the dilemma of vanishing gradient.

While there are several variants about the order of the units inside a residual block shown in Figure 3.16, the full pre-activation variant was adopted in this work due to better performance in reducing testing error and regularising capability to tackle overfitting, as stated by He et al. [2016]. The order of the full pre-activation variant inside the residual block is depicted in Figure 3.17 where: $input \rightarrow [Batch\ Norm \rightarrow ReLU \rightarrow weight] * 2 \rightarrow addition \rightarrow output$.

In this thesis, we mainly deal with volumetric medical images where an extended version of the original 2D ResNet by He et al. [2015] is implemented. A schematic

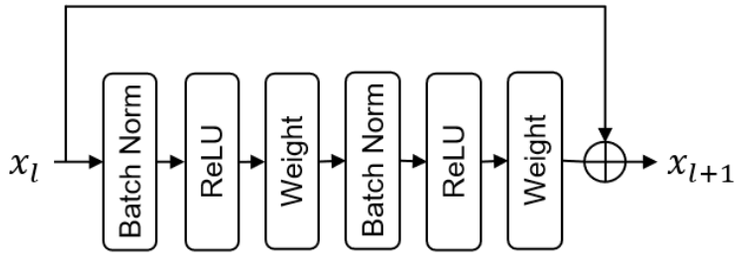


Figure 3.17: Full pre-activation order of the units inside residual block. Image reproduced from paper by He et al. [2016].

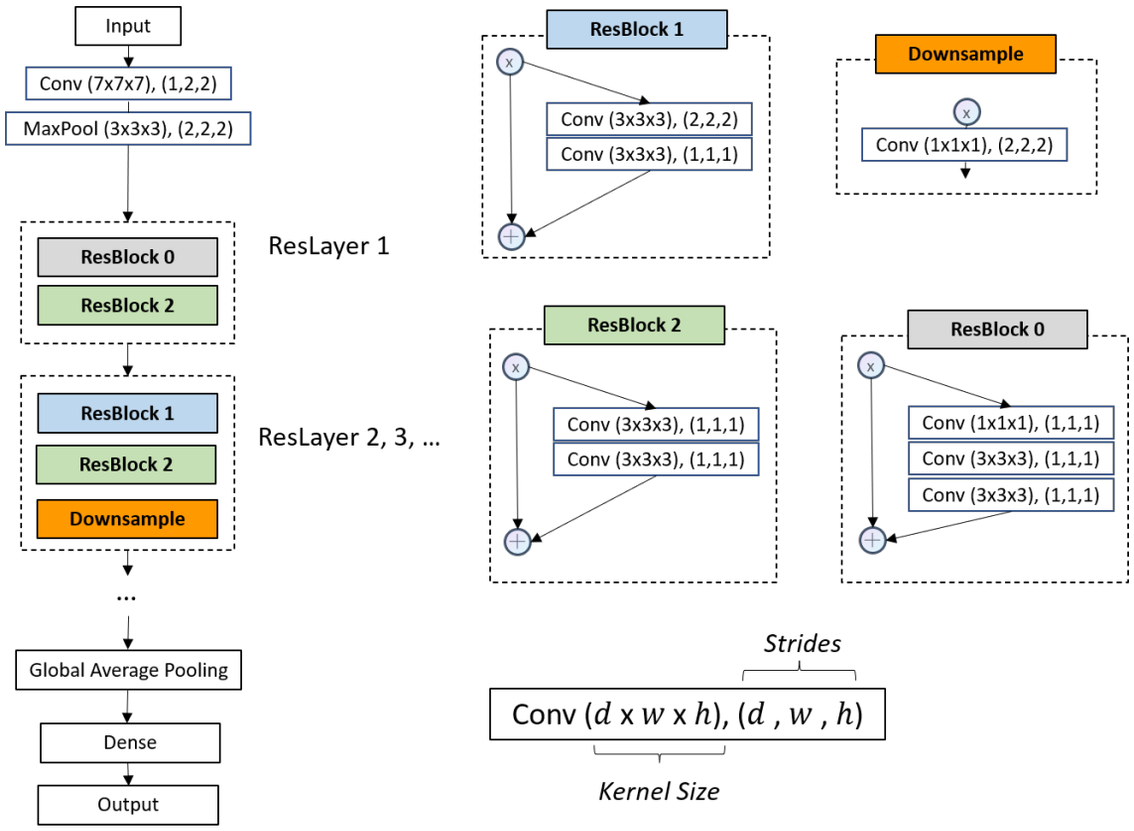


Figure 3.18: 3D-ResNet model architecture with ResLayer 1 differing from ResLayer (>2). ResNet model can be extended to deeper model by stacking ResLayer 2 like modules. All convolution operations and maxpooling operations are conducted in three dimensional settings.

plot of 3D ResNet model architect is illustrated in Figure 3.18. The depth of ResNet model can be extended by stacking a series of reslayer deeper into the network. The naming scheme of ResNet models often associate with the number of layers with trainable parameters, typically, ResNet18, ResNet32, ResNet50, etc.

3.3 Deep Metric Learning

In machine learning, classification is a task that predicts or assigns a class given a input. Typically, classification tasks in deep learning are often associated with softmax output and cross-entropy as a loss function, which in this work is known as

softmax loss. By implementing softmax loss the chosen model attempts to learn a mapping such that decision boundaries could be drawn to separate different classes. Such approach is inferior in metric learning where similarities or discrepancies of data representations are not enacted. Deep metric learning is an approach that attempts to optimise an embedding from deep neural networks utilising distance metrics such that similar objects are pulled towards each other and dissimilar objects are repelled against each other, as shown in Figure 3.19. In other words, metric learning aims to make objects in the embedding space discriminate enough given a margin rather than mere separability. In this work, we hypothesise that deep metric learning is able to empower discriminative feature learning for enhancement of prostate cancer classification among patients.

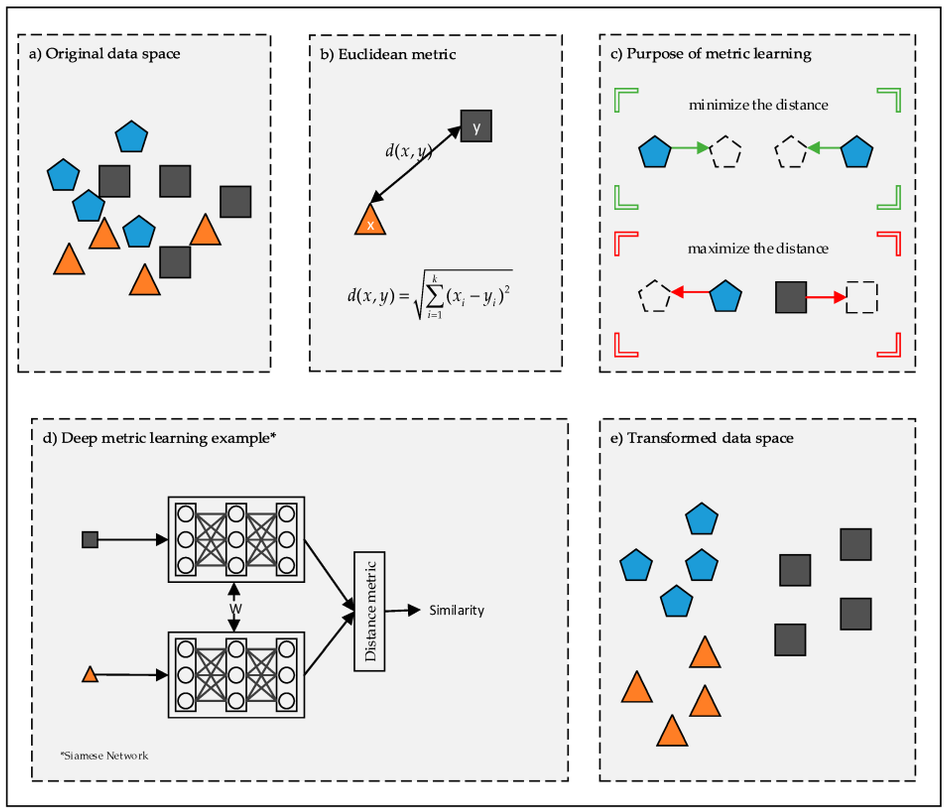


Figure 3.19: Objectives of deep metric learning. Image source from Kaya and Bilge [2019].

In machine learning, all metric learning algorithms are built on the basis of Mahalanobis distance metric (Kaya and Bilge [2019]). Mahalanobis distance between x and x' could be written in the form equation 3.15, where $L \in \mathbb{R}^{d \times N}$, such that L is a real value parameter matrix of shape *(number of dimensions, number of features)*.

$$D(x, x') = \sqrt{(Lx - Lx')^T (Lx - Lx')} \tag{3.15}$$

Mahalanobis distance satisfies the property of metric, typically, non-negativity, symmetry and triangle inequality. Kaya and Bilge [2019] stated that due to these properties, Mahalanobis distance in the original space is equal to the Euclidean distance in the transformed space. Generally, stated by Chan [2021], deep metric

learning can be categorised into two stages, mainly contrastive approaches and state-of-the-art approaches.

3.3.1 Contrastive Approaches

The core idea of contrastive approach correlates with a loss function that minimises the embedding distance of samples with the same label and maximises the embedding distance of samples with different labels. Chopra et al. [2005] proposed contrastive loss, as written in equation 3.16, where d is the distance metric, typically euclidean distance, $\mathbb{I}$ is the identity function equal to 1 for same label and 0 otherwise, α is denoted as the margin.

$$L_{contrastive} = \mathbb{I}_{y_i=y_j} d(x_i, x_j) + \mathbb{I}_{y_i \neq y_j} max(0, \alpha - d(x_i, x_j)) \tag{3.16}$$

While constrastive loss provides pairwise mining for distance optimisation, the capability of contrastive loss is constrained by pairwise objects while lacking reference point for the validity of positive pairs and negative pairs. In recent years, Schroff et al. [2015] proposed triplet loss that implements triplet mining for deep learning network, such that given a referencing sample, anchor (a), the model learns to minimise the distance between positive (p) pairs, i.e. the sample with the same label as the anchor, on the flip side, maximise the distance between negative (n) pairs, i.e. the sample with different labels from the anchor. A schematic diagram about the triplet loss is illustrated in Figure 3.20.

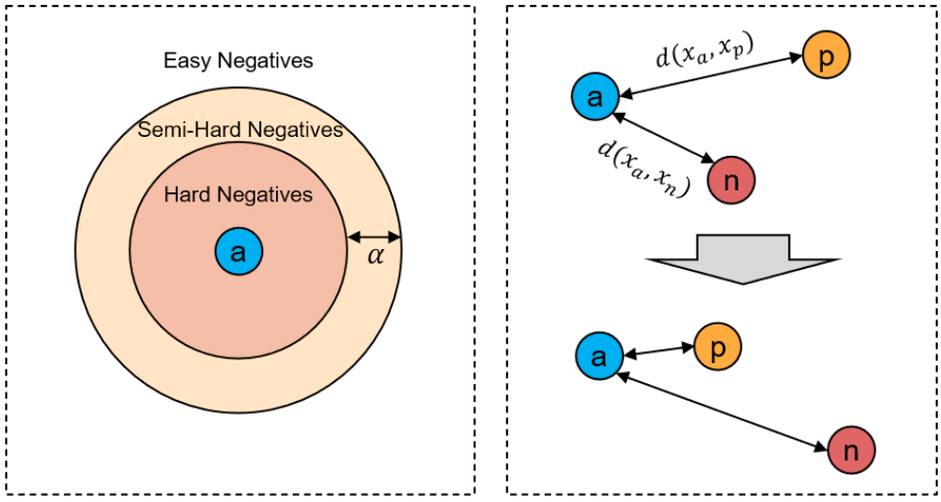


Figure 3.20: (Left) Criterion of negative pairs, where α is the margin parameter. (Right) Triplet loss attempts to minimise the distance between the positive and the anchor, $d(x_a, x_p)$, and maximise the distance between the anchor and negative, $d(x_a, x_n)$.

Triplet loss is expressed in equation 3.17, where α is the margin.

$$L_{triplet} = max(d(x_a, x_p) - d(x_a, x_n) + \alpha, 0) \tag{3.17}$$

Based on the negative criterion as shown in Figure 3.20, triplet mining can be defined as follows:

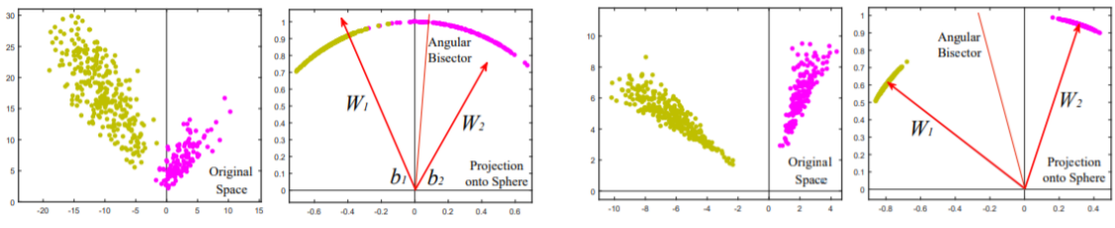
- **Easy Triplets**: Negative outside of the margin boundary, where $d(x_a, x_p) + \alpha < d(x_a, x_n)$
- **Semi-hard Triplets**: Triplets where the negative lies within the margin boundary, where $d(x_a, x_p) < d(x_a, x_n) < d(x_a, x_p) + \alpha$
- **Hard Triplets**: Triplets where negative is nearer to the anchor than a positive, $d(x_a, x_n) < d(x_a, x_p)$

While triplet loss performance is more superior to constrastive loss (Schroff et al. [2015], Hermans et al. [2017], Roth et al. [2020]), undeniably, contrastive approaches suffers from sampling or mining strategy. During training phase, the need to ensure the availability of triplets in the batch or in the embedding is crucial. Triplet mining could be done offline or online.

- **Offline Triplet Mining**: All the embeddings of the training set is computed and select hard or semi-hard triplets. The training is updated on these triplets. Generally, this method is not efficient with large amount of dataset such that a full pass run on the training set is obligated to generate triplets.
- **Online Triplet Mining**: Proposed by Schroff et al. [2015], triplets are generated on the fly for each batch. This technique is more efficient compared to offline mining, but does not perform well when the batch size is small where the number of samples per class is less than the number of classes.

In the paper by Hermans et al. [2017], the authors proposed batch all strategy for triplets mining such that the hardest positive, i.e. the positive furthers from the anchor, and the hardest negative, i.e. the negative closest to the anchor is selected from the batch. However, this would lead to class collapse (Levi et al. [2020]), a phenomenon in the embeddings where the model fails to distinguish between positive and negative pairs, such that $d(x_a, x_p) = d(x_a, x_n)$ and equation 3.17 returns only the value of the margin, α.

3.3.2 Margin-based Softmax Loss Approaches



(a) Original Softmax Loss **(b)** A-Softmax Loss

Figure 3.21: Intrinsic angular properties of softmax loss in the original space. Yellow dots and purple dots represent different class features respectively. W_1 and W_2 act as class center. Image source from Liu et al. [2017c].

Contrastive approaches only aim to optimise an embedding but does not map an embedding to class labels for classification task. Therefore, contrastive approaches

rely on pseudometric such as k-NN to implement classification while hampering the model to learn the objective end-to-end. Furthermore, the requirement of carefully handcrafted sampling strategies and class collapse associated from constrastive approaches is not trivial and shift researchers to take a step back on softmax approach with margin criterion, known as margin-based softmax loss. Researchers like Liu et al. [2017c] question the appropriation of euclidean margin to maximise inter-class distance, at the same time, minimise intra-class distance. It is verified by Wen et al. [2016] that features learnt by softmax loss possess angular properties as shown in Figure 3.21. Liu et al. [2017c] proposed angular softmax (A-softmax) loss that projects features on a hypersphere and optimise the geodesic distance of the classes through softmax loss. A-softmax loss is derived from softmax loss as shown in equation 3.18, where the weights and features are normalised, $||W_i|| = 1, ||x_i||$, and taking the bias as zero, $b_i = 0$. The equation could be rewritten as equation 3.19, where $cos(\theta)$ is the cosine distance between $\mathbf{W}$ and $\mathbf{x}$. Since $||x||$ does not contribute to the cost function, we can fix $||x|| = s$ where s is a hyperparameter that define the radius of the sphere (Deng et al. [2018]) and scale the logits of the model. On the other hand, θ controls the separability of the decision boundary and by imposing multiplicative margin (m), such that $cos(m\theta_i) > cos(\theta_i)$, A-softmax loss could encourage intra-class compactness as shown in Figure 3.21. A-softmax loss is reduced and expressed in equation 3.20.

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(\mathbf{W}_{y_i}^T\mathbf{x}_i + b_{y_i})}{exp(\mathbf{W}_{y_i}^T\mathbf{x}_i + b_{y_i}) + \sum_{j\neq y_i} exp(\mathbf{W}_j^T\mathbf{x}_i + b_j)} \tag{3.18}$$

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(||\mathbf{W}_{y_i}^T||||\mathbf{x}_i||cos(\theta_{y_i}))}{exp(||\mathbf{W}_{y_i}^T||||\mathbf{x}_i||cos(\theta_{y_i})) + \sum_{j\neq y_i} exp(||\mathbf{W}_j^T||||\mathbf{x}_i||cos(\theta_j))} \tag{3.19}$$

$$L_{asoftmax} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(s\,cos(m\,\theta_{y_i}))}{exp(s\,cos(m\,\theta_{y_i})) + \sum_{j\neq y_i} exp(s\,cos(\theta_j))} \tag{3.20}$$

The intervention of attempting to optimise features geodesic distance in the hypersphere with angular margin using asoftmax loss has spark a series of novel methods, most notably cosface loss by Wang et al. [2018] and arcface loss by Deng et al. [2018].

Cosface loss is expressed in equation 3.21 through additive margin on cosine distance.

$$L_{cosface} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(s\,cos(\theta_{y_i}) + m)}{exp(s\,cos(\theta_{y_i}) + m) + \sum_{j\neq y_i} exp(s\,cos(\theta_j))} \tag{3.21}$$

Arcface loss is written as equation 3.22 with additive penalty on the angle dictating the cosine distance.

$$L_{arcface} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(s\,cos(\theta_{y_i} + m))}{exp(s\,cos(\theta_{y_i} + m)) + \sum_{j\neq y_i} exp(s\,cos(\theta_j))} \tag{3.22}$$

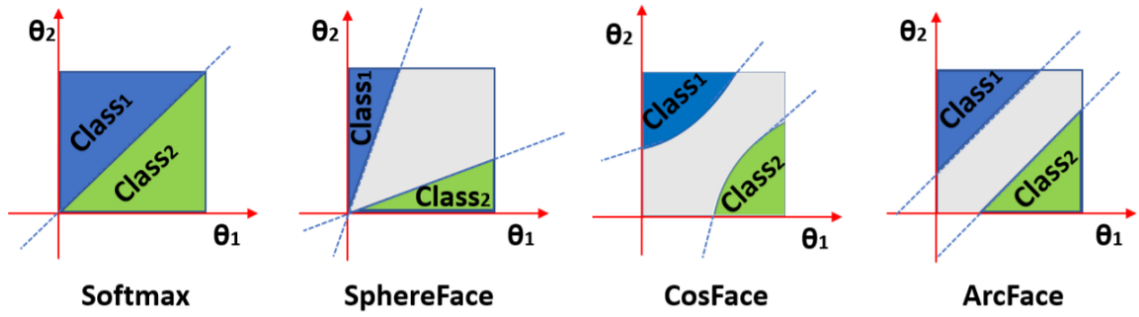
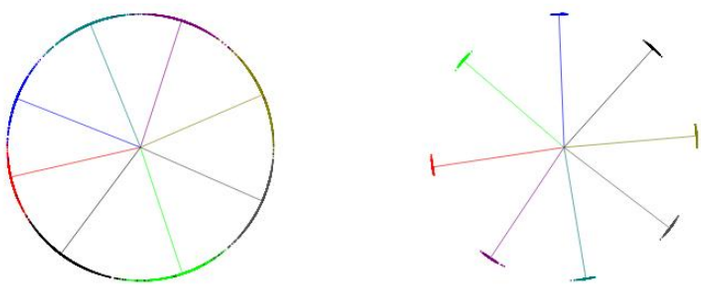


Figure 3.22: Decision boundaries of softmax loss, Sphereface loss(asoftmax loss), Cosface loss and Arcface loss. Image source by Deng et al. [2018].

The implementation of Cosface and Arcface loss is similar but ablation studies by various papers (Deng et al. [2018], Wu et al. [2017], Roth et al. [2020]) show slightly better performance in use cases deriving from Arcface loss. The decision boundaries of softmax, asoftmax, Cosface and Arcface loss are illustrated in Figure 3.22. Because Arcface loss directly penalises the angle via additive angular margin, it could maintain a linear angular margin within the interval. A proof of concept for Arcface loss enforcing compactness of similar class with respect to conventional softmax loss is depicted in Figure 3.23.



(a) Softmax (b) ArcFace

Figure 3.23: Intra and iter-class separability of (a) softmax loss vs (b) Arcfaceloss. Image source by Deng et al. [2018].

3.3.3 Content Based Image Retrieval (CBIR)

CBIR is a method that utilised image as a query to retrieve information from image data set. In medical applications, CBIR could effectively help to improve early diagnosis via visual aids through the retrieval of semantically similar images from the database associated with supporting information, such as, bio-markers, electronic health records (EHR), patient id, etc. In the work by Zhong et al. [2021], they emphasised that the main property for an accurate and faster retrieval of a CBIR system relies on the image embeddings. Image embeddings are mapping of low dimensional representations that consisted of encoded information from the original image space. These lightweight image embeddings could formulate a feature database such that efficient and accelerated CBIR is made possible as compared to original images.

While CBIR task is not the main objective of this work, we attempt to demonstrate that the optimisation of an embedding space for discriminative image representation with deep metric learning could enhance retrieval task capability. In such a way, doctors could benefit from the retrieved image empowered by clinically meaningful distance metric to support decision making. It is worth noting that we did not include supporting patient information other than images in this work. The CBIR capability of this thesis would only be an image to image retrieval task.

4. Implementation

This chapter discusses the details of data set for this work as well as data preparation pipeline for the model. Next, the proposed methods would be discussed following the hyperparameters and evaluation metrics.

4.1 Data Set

In this thesis, we collaborate with Alta Klinik for prostate diseases medical image analysis, where we received mpMRI images from Alta Klinik. In total, we received MRI images in .NRRD format from 1873 patients. Table 4.1 shows the range of image size (number of pixels in each dimension) and image spacing (distance between pixels) of MRI images of different MRI sequences (T2, DWI, ADC).

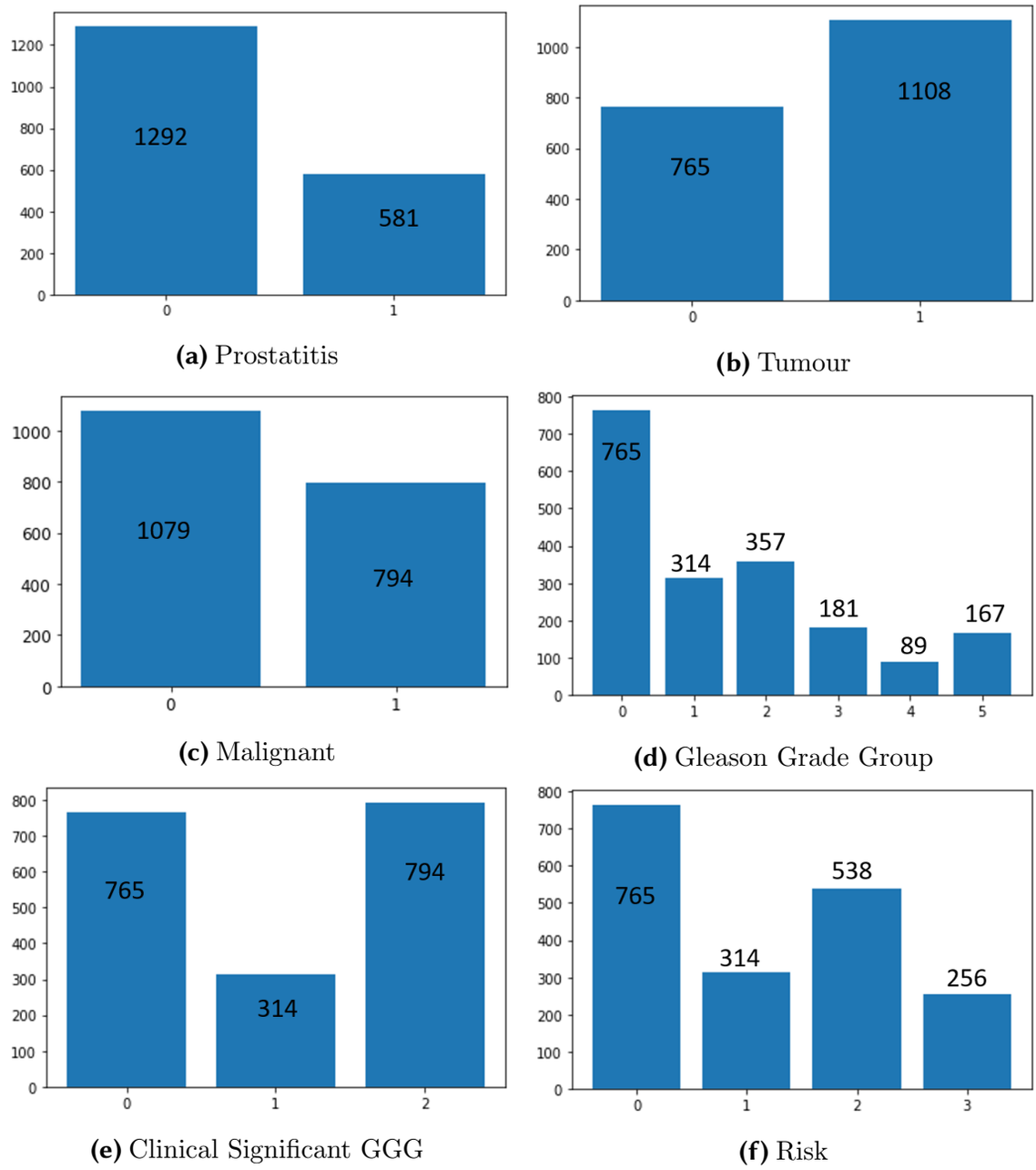
	Range	**Depth (D)**	**Width (W)**	**Height (H)**	**Spacing (D, W, H)**
T2	Min.	13	200	256	(0.54, 0.54, 3.89)
	Max.	48	896	1024	
DWI	Min.	13	128	128	(1.67, 1.67, 3.00)
	Max.	45	276	296	
ADC	Min.	13	128	128	(1.67, 1.67, 3.00)
	Max.	74	524	524	

Table 4.1: Properties of MRI data sets from different MRI sequences (T2, DWI, ADC) with depth, width, height and spacing.

Alta Klinik also provide us tabular information about features of each patients, including presence of tumour, Gleason group grade (GGG) and presence of prostatitis. The aggregation of data set could be listed as the follow granularity according to its criterion.

- **Prostatitis**: Patients with or without prostatitis
- **Tumour**: Presence of tumour in the patient prostate

- **Malignant**: Patient with tumour and GGG>1
- **GGG**: Gleason grade score from 1 to 5
- **Risk**: Patients without Gleason grading as no risk; Low, medium and high risk according to GGG
- **Clinical Significant GGG**: Patients without GGG; Patients with GGG1; Patients with GGG>1



(a) Prostatitis **(b)** Tumour

(c) Malignant **(d)** Gleason Grade Group

(e) Clinical Significant GGG **(f)** Risk

Figure 4.1: Distribution of prostatitis, tumour, malignant, GGG, clinical significant GGG and risk patients with the number of patients presented as well as corresponding labels listed in Table 4.2.

Distributions of data set granularity are depicted in Figure 4.1 as well as the label verbosity of each granularity listed in Table 4.2. In this work, the primary interest

lies on the ability of proposed method to classify malignancy of patients. The rest of the granularity prostatitis, tumour, GGG, clinical significant GGG and risk could be used as supportive inputs for decision boundary of patients with malignant prostate tumour, which we will further discuss in section 4.4.

	Granularity	**Label Verbose**
(a)	Prostatitis	{0: Prostatitis patient}, {1: No-prostatitis patient}
(b)	Tumour	{0: Tumour Patients}, {1: Non-tumour Patients}
(c)	Malignant	{0: Cancerous Patients}, {1: Non-cancerous Patients}
(d)	Gleason Grade Group	{0: No GGG}, {1: GGG1}, {2: GGG2}, {3: GGG3}, {4: GGG4}, {5: GGG5}
(e)	Clinical Significant GGG	{0: No GGG}, {1: GGG1 Patients}, {2: GGG>1 Patients}
(f)	Risk	{0: No risk or no GGG}, {1: Low Risk}, {2: Mid Risk}, {3: High Risk}

Table 4.2: Label distributions of corresponding granularity shown in Figure 4.1

4.2 Objectives

Illustrated in Figure 4.2, the main objective of this work is to draw a decision boundary to separate the patients with malignant prostate tumour, such that GGG is greater than one, from patients with non-malignant prostate tumour. This process is crucial to support biopsies decision such that false biopsies upon patients could be mitigated, preventing over-diagnosis.

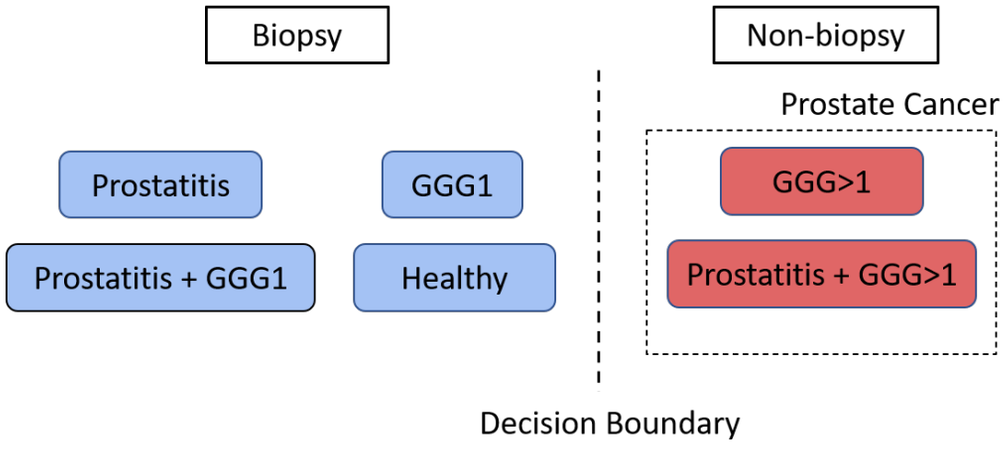


Figure 4.2: Decision boundary for "to biopsy or not biopsy?"

A schematic Venn diagram representing the objective boundaries of biopsy objective is depicted in Figure 4.3. In this context, healthy criteria represents patients that do not possess prostate diseases, viz. tumour-free and prostatitis not detected.

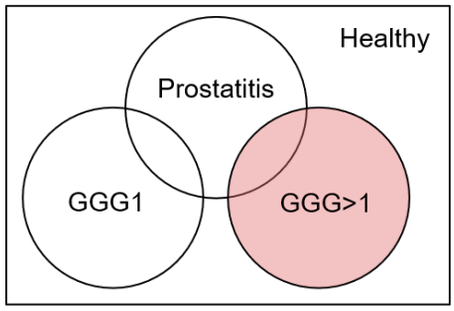


Figure 4.3: Overview of Venn diagram representing the various sets of patient attributes in the data set. The set shaded in red belongs to cancerous patients where biopsies should be assigned. A healthy set refers to patients that are not assigned with GGG and prostatitis not detected.

In this thesis, we proposed a deep metric learning technique, ArcFace loss, which encourages intra-class compactness and further inter-class separability. We hypothesise that the objective of ArcFace loss enacts discrimination of features from different classes given a margin such that separability of the decision boundary for biopsy could be more distinctive. We also attempt to exploit ArcFace loss on volumetric images with different MRI sequences (T2, DWI, ADC) for prostate cancer classification. As a complimentary task of this thesis, we attempt to extend the usage of visual similarity in medical imaging. We hypothesise the clinically meaningful distance metric learned by ArcFace loss may serve as a measurement for critical clinic decision support. Concurrently, we aim to exploit the capabilities of content based image retrieval (CBIR) on 3D images through visual similarity.

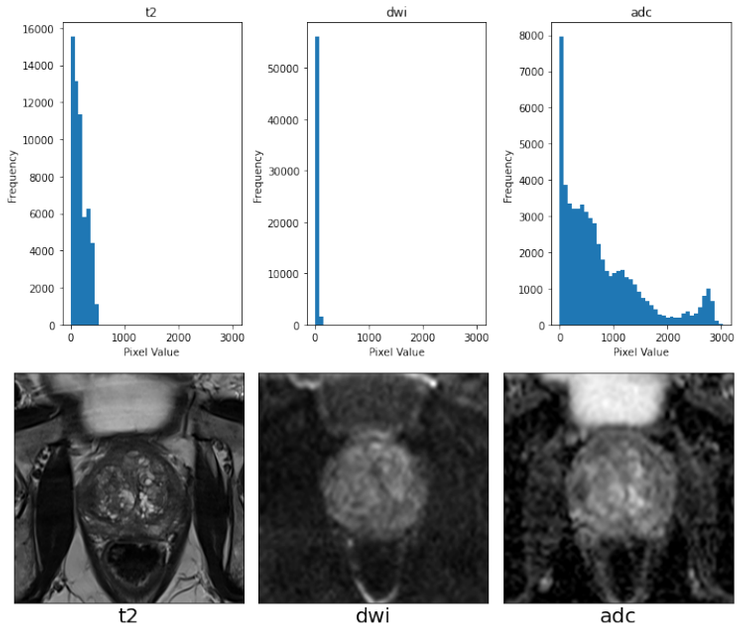
4.3 Data Preprocessing

In section 4.1, it is observed that the size and spacing of various MRI sequences are not homogeneous. To streamline the data pipeline for our model, homogeneity of dimensional representations is crucial, where target dimensions are listed in table 4.3. We first resize the volumetric data to achieve the shape of (32, 250, 250) and normalise the image spacing to (3.0, 3.0, 0.5). For more efficient dataloading, we concatenate the MRI sequences along the channel axis forming dimensional representations of (32, 250, 250, 3). Computational resources could be further mitigated by centre cropping the images into patch size of (32, 144, 144, 3).

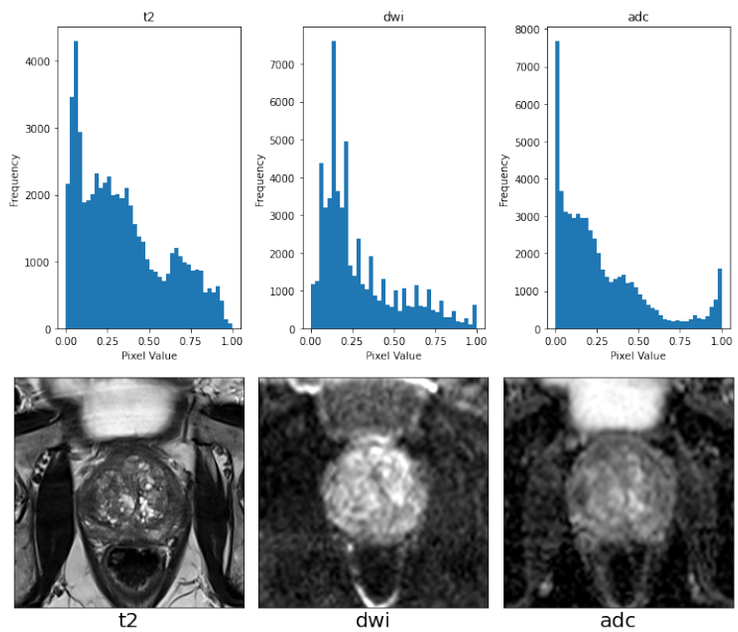
	Dimensions
New size	(32, 250, 250)
New spacing	(3.0, 0.5, 0.5)
Image shape	(32, 250, 250, 3)
Patch size	(24, 144, 144, 3)

Table 4.3: Dimensional properties of volumetric images where a tuple of 3 values, corresponds to dimesional representation of *(depth, width, height)*, whereas, a tuple of 4 values, corresponds to *(depth, width, height, channel)*.

In deep learning, it is common practice to standardise and normalise the image intensity of each MRI sequence respectively since the image intensity of MRI images



(a) Original Image Plot



(b) Preprocessed Image Plot

Figure 4.4: Comparative plot of original and preprocessed image of a prostate cancer patient from the mid slice. (Top) Histogram of pixel intensities, with it's corresponding (Bottom) image plot for different MRI sequences given both (a) original data set and (b) preprocessed data set.

consists of high intensity values (typically up to thousands) that would hinder the performance of deep learning models. Standardisation would centralise the intensity range, while min-max normalisation would reduce the intensity range in between 0 to 1. In this work, we first purge outlier pixels by clipping the image intensity between 1^{st} to 99^{th} percentile. Next, the channel-wise standardisation is computed with equation 4.1 through the subtraction of mean values along the channel axis and

of unit variance channel wise. After standardisation, the images are normalised with channel wise min-max-normalisation as denoted in equation 4.2.

$$x_{std} = \frac{x - \mu}{\sigma} \tag{4.1}$$

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)} \tag{4.2}$$

Comparative plot of original images and preprocessed images for respective MRI sequences are duly plotted in Figure 4.4.

4.4 Model Architecture

4.4.1 Backbone Model

In section 3.2.4 we introduced the building blocks of 3D-ResNet centering around the idea of skip connections for reduced trainable parameters to prevent vanishing gradient for effective training. However 3D-ResNet assumes a cubic like volumetric input or balanced depth to width/height ratio, such property is known as isotropic. Height/width resolution preserved comparatively more information than depth. Such settings contribute to the rapid diminishing of dimension along the depth axis during the downsampling process. In this thesis, the input MRI data has low depth to width/height ratio such that conventional 3D ResNet model is not able to benefit feature extractions for anisotropic settings particularly along the depth dimension. Therefore we adapted Anisotropic Hybrid ResNet3D (AH-ResNet) model proposed by Liu et al. [2017b] to tackle anisotropic properties of volumetric images.

Figure 4.5 depicts schematic graph about AH-ResNet10 model architecture. We did parameter sweeps for the model depth with AH-ResNet10, 18, 32 and 50 by increasing reslayer and model width with initial filters of 64, 32, 16 and 8. We concluded that AH-ResNet10 with initial filters of 64 provides the best results and is least subjected to overfitting. The features after global average pooling layer is directly connected to an output layer (head) instead of passing through a fully connected layer. This design attempts to reduce the complexity of the model for mitigation of overfitting phenomenon. Furthermore, there is a dropout layer prior to global average pooling layer and batch normalisation layer after global average pooling layer for regularisation effects.

4.4.2 Arcface Head

Figure 4.6 illustrate the computation process of Arcface loss with complimentary pseudo-code summarised in Algorithm 1. Features from the backbone model, x_i and the weights, W_j are normalised and multiplied to obtain cosine similarity, $cos\theta_j$ (logits). Angle between feature and ground truth weight, θ_{y_i} for a given ground truth is retrieved through $arccos(cos\theta_{y_i})$. W_j could be interpreted as centre for each classes. Additive margin penalty, m is added on target angle θ_{y_i}. Taking cosine on target logits, yield $cos(\theta_{y_i} + m)$. The logits are scaled accordingly with logits scale,

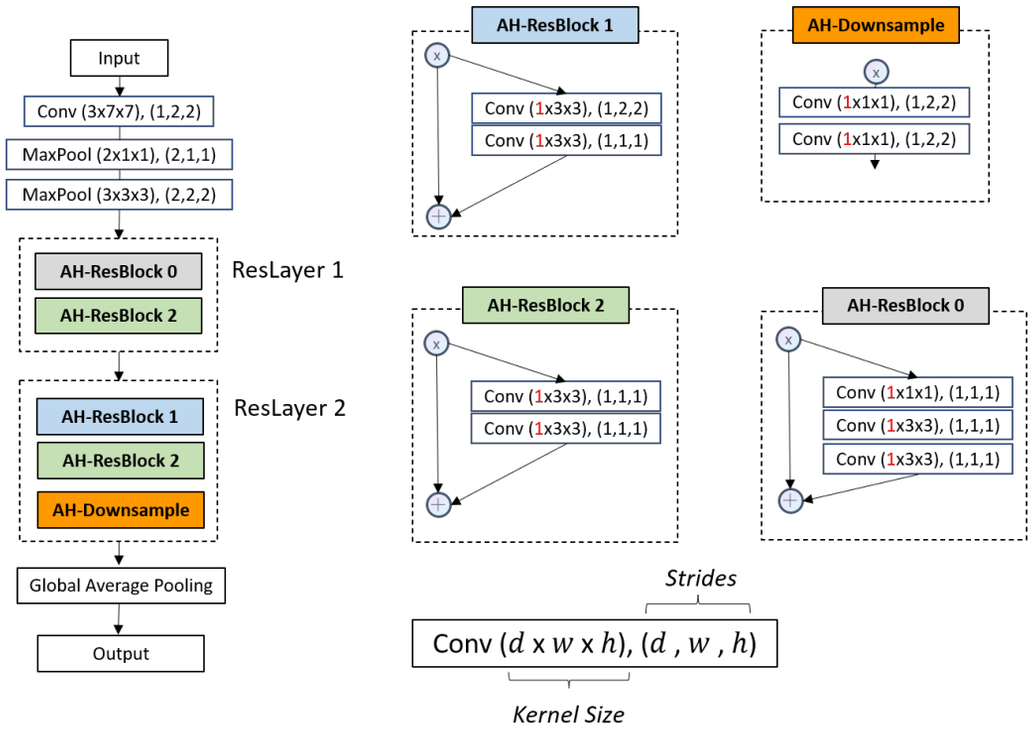


Figure 4.5: AH-ResNet10 with reduced downsampling along the depth axis for pooling and convolution strides. Kernel size of Conv3D is reduced to 1 *(red)*.

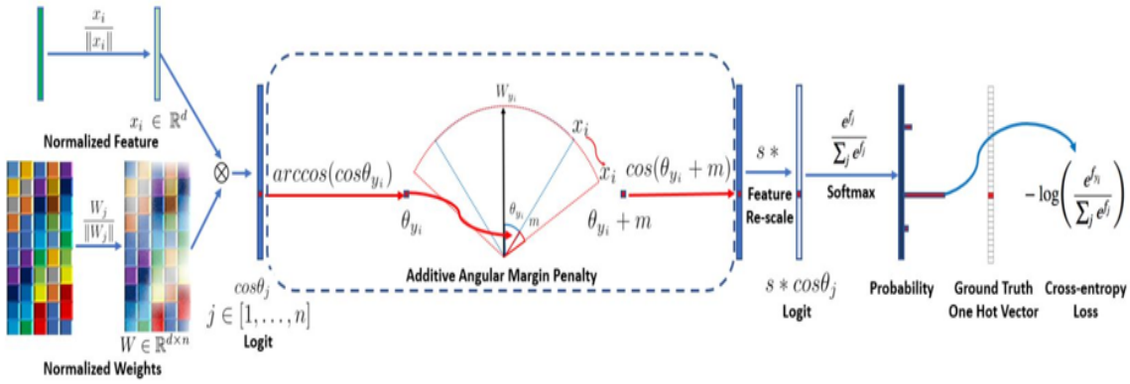


Figure 4.6: Schematic diagram about algorithm of Arcface loss. Image source from Deng et al. [2018].

s. Mapping the logits to probability through softmax function and computing cross entropy loss are then done as expressed in equation 4.3.

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(s\,cos(\theta_{y_i} + m))}{exp(s\,cos(\theta_{y_i} + m)) + \sum_{j \neq y_i} exp(s\,cos(\theta_j))} \tag{4.3}$$

For clarity, in this work, Arcface head refers to output layer paired with Arcface loss.

Unlike conventional softmax loss, the logits are linear functions where the value range from $[-\infty, \infty]$ which could map the softmax output to $[0, 1]$. Whereas for Arcface loss, the logits are derived from cosine mapping, where the range for $cos\theta$ range from

Algorithm 1: Arcface Pseudo-code with Tensorflow

Input : Logits scale s, Angular margin(in radian) m, Number of classes n, Groundtruth gt
Output : Scaled Logits

```
/* Initialise weights */
W ← add_weight(shape = (input.shape[1], n))
/* Operations */
x = tf.nn.l2_normalize(x, axis = 1)
W = tf.nn.l2_normalize(W, axis = 0)
cos_theta = tf.matmul(x, W)
cos_theta = tf.clip_by_value(cos_theta, −1.0 + 1e − 7, 1.0 − 1e − 7)
if training then
    mask = tf.one_hot(gt, depth = n)        // position of ground truth
    theta = tf.acos(cos_theta)
    cos_mt = cos(theta + m)                  // additive angular margin
    logits = tf.where(mask == 1, cos_mt, cos_theta)
else
    logits = cos_theta          // no margin penalty during inferencing
end if
logits = tf.multiply(logits, s)
return logits
```

$[-1, 1]$. The cosine mapping as logits are not able to map the softmax output to $[0, 1]$. Therefore, logits scale s attempts to scale the logits such that the probability output from a softmax function with Arcfaceloss could achieve high confidence bounded by $[0, 1]$. In research papers inspired by angular softmax (Liu et al. [2017c], Wang et al. [2018], Deng et al. [2018]), logits scale s, is empirical derived through parameter tuning. Wang et al. [2018] provides an approximation for lower bound of s, expressed in equation 4.4. Where P_w is the expected minimum posterior probability of class center, W and C denotes the number of classes in the data set.

$$s \geq \frac{C-1}{C} log \frac{(C-1)P_w}{1-P_w} \tag{4.4}$$

Zhang et al. [2019] extends the work by introduction a better approximation, shown in equation 4.5.

$$s \approx \sqrt{2} log(C-1) \tag{4.5}$$

However, both approximated s values are derived from Cosface loss. Hence, in this thesis, an approximated s lower bound is expressed in equation 4.6. The complete derivation of the lower bound is available on Appendix A.1.

$$s \geq -\frac{1}{2} log \left(\frac{1-P_w}{P_w(C-1)} \right) \tag{4.6}$$

4.4.3 Dense Head as a Baseline

Arcface head is a deep metric learning techniques that attempts to optimise the intra-class and inter-class geodesic distance through angular margin in the hypersphere.

To validate the credibility of our proposed method, the baseline implemented as a benchmark is the vanilla classification dense softmax output. Here, we refer dense head as the output layer paired with softmax loss. The output layer consisted of dense layer with units equivalent to the number of classes. The logits is then bypassed through a softmax function and compute cross entropy loss, as shown in equation 4.7.

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(\mathbf{W}_{y_i}^T\mathbf{x}_i + b_{y_i})}{exp(\mathbf{W}_{y_i}^T\mathbf{x}_i + b_{y_i}) + \sum_{j\neq y_i} exp(\mathbf{W}_j^T\mathbf{x}_i + b_j)} \tag{4.7}$$

4.4.4 Late Fusion

Three prostate MRI image sequences (T2, DWI, ADC) are provided in the data set. While it is common practice to liaise the image sequences as channel input for richer data features, however, it is not possible for all three image sequences to have the same alignment due to constraints. We hypothesise that each image sequence would contribute to different representations in the embedding space, by opting to feed the network on each image sequence respectively. The embedding output of three image sequences is concatenated forming an ensemble embedding that would enrich the embedding space. This form a late fusion where the concatenated embedding is connected to a fully connected network with a dimension size of 128 associating with a dropout layer before connecting to output layer, as shown in Figure 4.7.

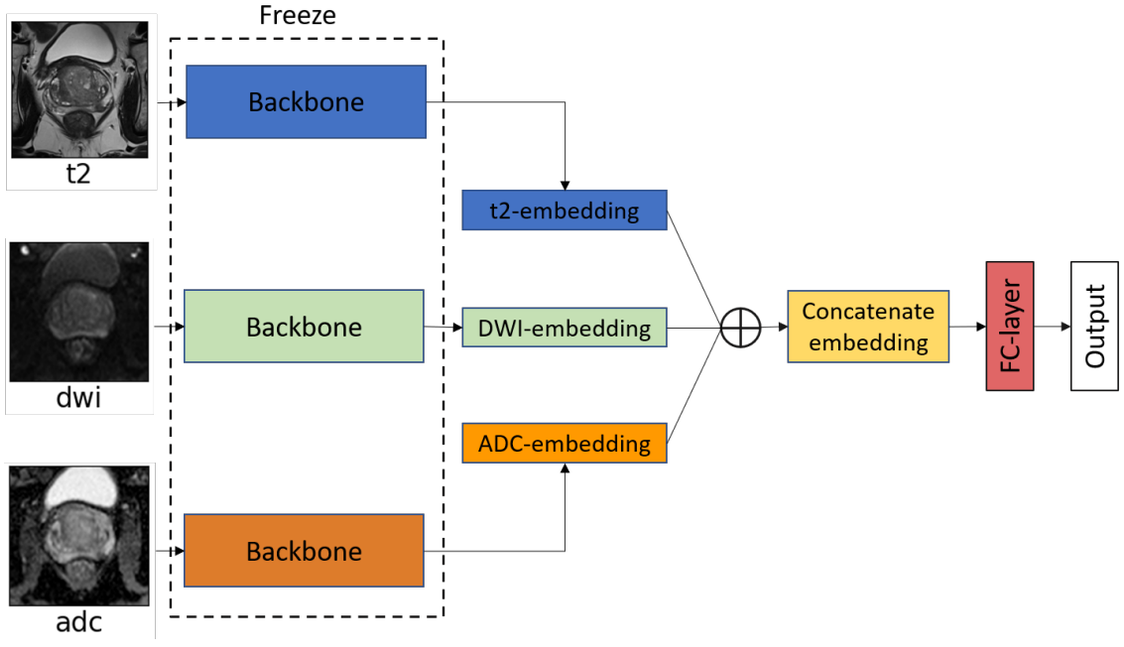


Figure 4.7: T2, DWI, ADC are trained on different models respectively where the models are freeze and the extracted embeddings from each image sequences are concatenated to form a late fusion. The concatenated embedding are connected to a fully connect network where we fine tuned this layer for better representation.

4.4.5 Head Settings

Generally, a network model would deal with one head for a single task in the downstream section, as depicted in Figure 4.8.

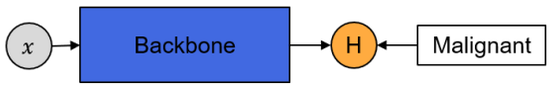


Figure 4.8: Single head model. Head, H could be dense head or Arcface head.

In this work, we hypothesise that by adding different granularity of the patient profiles, introduced in Table 4.2, we form a multi-head setting for a model such that each granularity would provide a decision boundary with respect to their objectives, simultaneously enriching the representations in the embedding space. In this work, the main focus remains on drawing a decision boundary between malignant and non malignant patients, where the remaining head could be deemed as regularisers for the model in the downstream task. A schematic figure illustrating multi head setting is shown in Figure 4.9.

Furthermore, each head of the multi-head setting is able to focus on their respective classification task where the latent space should figure out the features for optimal learning. This approach would prevent explosive combinations of classes (refer to Appendix A.6) based on the data set granularity mentioned in section 4.1. This mitigates the need to tackling each combinations of classes, while able to optimise the latent space at the same time.

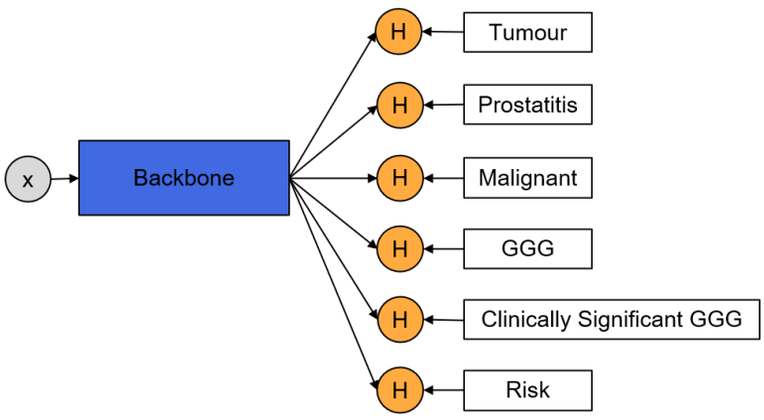


Figure 4.9: Multi head model. Head, H could be dense head or Arcface head.

4.5 Training and Inferencing

Validation of a deep learning predictions are performed on unseen data for model performance assessment. It is common practice to split the data set into training data and validation data. However, there is no guarantee that if a model performs well only on this particular split, it will explain the generalisation capability of a model. Therefore, we used k-fold cross-validation, a re-sampling technique that rotates through the k-folds for training and validation data. To prevent data leakage

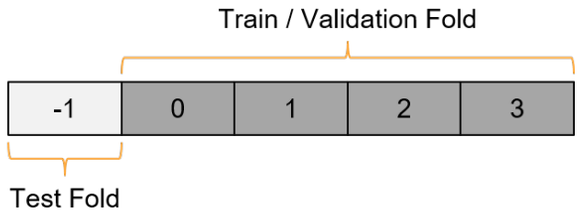


Figure 4.10: Training, validation and test data split. Test data possessed the index of -1, while training and validation data is formulated through the permutation of 0, 1, 2, 3 combinations.

in our data set, we also opt testing data in our work, which is the data independent of training and validation data, in other words, the true unseen data.

In this thesis, the data set is split into stratified 5 folds, where the distribution of labels for each folds are the same, as shown in Figure 4.10. Each fold consisted of approximately 375 patients. The first fold is used as the test set, whereas the following 4 folds are used as training and validation set.

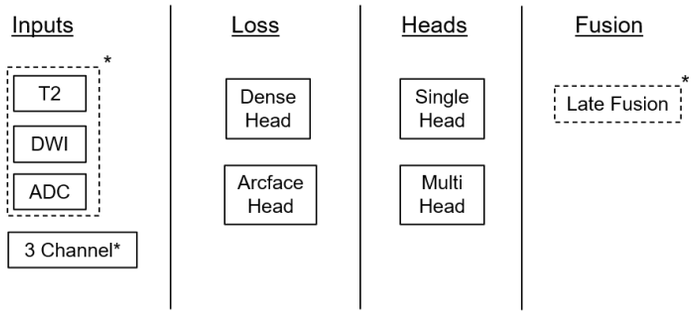


Figure 4.11: Inputs, losses, head settings and fusion combinations for ablation studies. 3 channel refers to T2+DWI+ADC as channel inputs. *Late fusion is only available for model with individual image sequences as input.

Prior to late fusion, this work serves to investigate the impact of image sequences as an input given a model performance. Figure 4.11 illustrates the possible combinations of ablation studies that could be carried out. In total, the combination formulates in total 20 ablation studies for a training data set. With 4 fold cross validation, in total 80 runs would be conducted for the investigation of this thesis.

4.5.1 Training Phase

Data augmentation is implemented in the training phase to tackle overfitting with the transformations and value range listed in Table 4.4. The optimiser used during model training is Adam by Kingma and Ba [2014].

Shown in Figure 4.2, the distribution of the classes in the data set are imbalanced. A common strategy to tackle class imbalanced is to implement weighted cross entropy. Where the weights, $w_t \in [0, 1]$ are the normalised inverse class frequency, shown in equation 4.8.

$$L = -w_t \, log(p_t) \tag{4.8}$$

Transformation	Value Range
Random Crop	Width: 166, Height: 166
Elastic Deformation	Shear stress σ: [20, 40]
Rotation	z-axis: ± 24
Scale Transformation	± 1.15
Mirror Transformation	y-axis
Brightness	[0.7, 1.5]
Gaussian Noise	μ=0.0, σ=0.5
Gaussian Blur	μ=0.5, σ=2.0
Contrast	[0.75, 1.25]

Table 4.4: Transformations and value ranges of data augmentations.

Lin et al. [2017] address the problem of cross entropy where easy classified examples contribute to the loss. A modulating factor, $-(1-p_t)^{\gamma}$, was introduced by Lin et al. [2017] to penalise easy examples (see Figure 4.12), known as focal loss, expressed in equation 4.9, where $\gamma \in [0, 5]$. We adapted weighted focal loss in our work to address the issue of class imbalance.

$$FL = -w_t(1-p_t)^{\gamma} log(p_t) \tag{4.9}$$

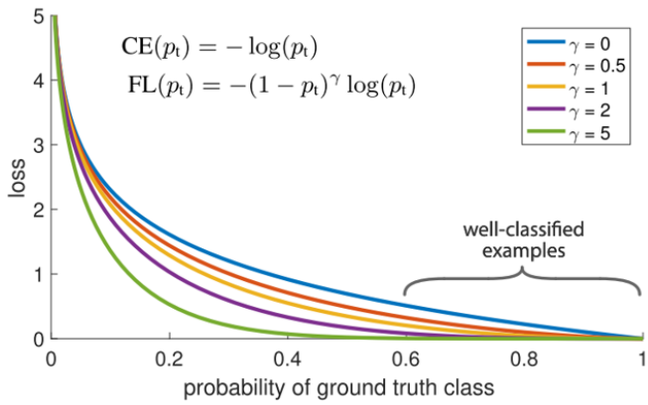


Figure 4.12: Comparison of loss for cross entropy and focal loss with γ value ranging from $[0, 5]$. When $\gamma = 0$, focal loss becomes cross entropy loss. For $\gamma > 0$, the modulating factor penalise the relative loss for easy classified examples, where $p_t > 0.5$ and focus only misclassified examples. Image source from Lin et al. [2017].

4.5.2 Inference Phase

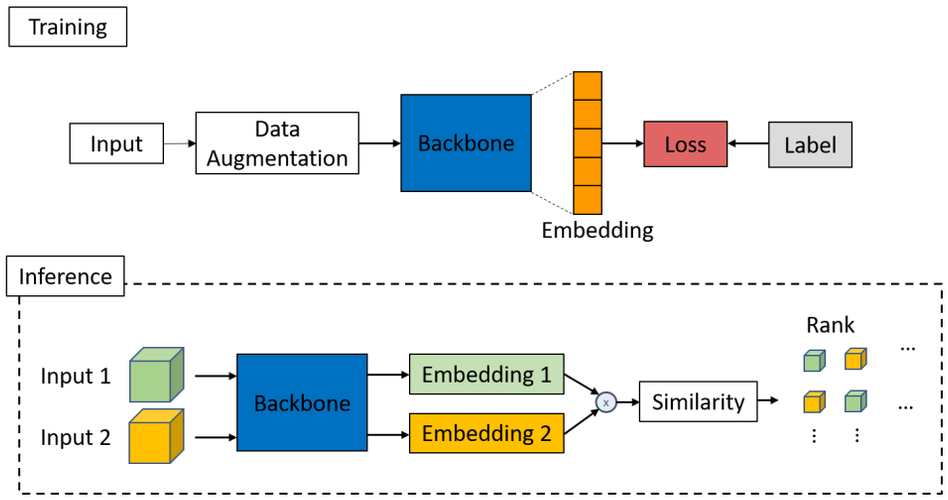


Figure 4.13: Implementations of training and inferencing phase.

During the inference phase, data augmentation technique is not implemented. In addition to the classification task, in the inference phase, the similarity of each input are computed through the multiplication of each embedding input respectively to achieve a cosine similarity. The scores are then rank in a descending order with the input image as the anchor whilst the neighbour objects with the highest score being close to the anchor. The differences between training phase and inferencing phase is shown in Figure 4.13. At this stage, CBIR is performed with the consent of similarity scores ranking in the nearest neighbourhood with input image as a query and the image embeddings of the validation data set as the feature database.

4.6 Evaluation Metrics

Metrics are measurements that assess the performance of the proposed model.

4.6.1 Classification Evaluation Metrics

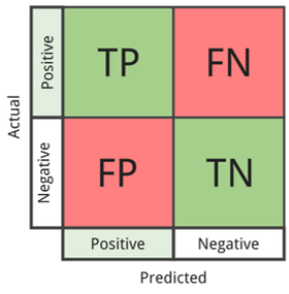


Figure 4.14: Confusion matrix where, True Positive (TP): Label and prediction are both positive. False Negative (FN): Label is positive, but prediction is negative. True Negative (TN): Label and prediction are both negatives. False Positive (FP): Label is negative but prediction is positive. Image source by MLee [2021].

- **Precision**: Ratio of true positives over all positives.

$$\frac{TP}{TP + FP} \tag{4.10}$$

- **Recall**: Measurement of the model correctly classifying true positives.

$$\frac{TP}{TP + FN} \tag{4.11}$$

- **F1-Score**: Harmonic mean of precision and recall.

$$\frac{2TP}{2TP + FP + FN} \tag{4.12}$$

- **Accuracy**: Ratio of total number of predictions over total number of predictions.

$$\frac{TP + TN}{TP + TN + FP + TN} \tag{4.13}$$

- **Area under the ROC Curve(AUC)**: Receiver operating characteristic(ROC) curve is a graph that plots the trade off between true positive rate(TPR) and false positive rate(FPR). AUC is the area under the curve where it calculates the aggregated performance of a classification model for every possible classification thresholds (Google [2020]).

4.6.2 Retrieval Evaluation Metrics

- **Precision@k (P@k)**: Ratio of matched queries in top-k neighbour over top-k neighbour.

$$\frac{\text{True Positive at top-k}}{\text{top-k}} \tag{4.14}$$

- **Recall@k (R@k)**: Measurement of queries having at least one neighbour retrieved in the first k result (Musgrave et al. [2020]).
- **R-Precision**: Defined as finding the R nearest references to the query (Musgrave et al. [2020]).

$$\frac{\text{True Positive at R}}{\text{R}} \tag{4.15}$$

- **Mean Average Precision (MAP)@R**: Mean of precision in the nearest neighbour for R sample set.

$$\frac{1}{R}\sum_{i=1}^{R} P(i) \text{ , where } P(i) = \begin{cases} \text{precision at i,} & \text{if the i-th retrieval is correct} \\ 0, & \text{otherwise} \end{cases} \tag{4.16}$$

4.7 Hyperparameters

Hyperparameters are parameters that contribute to the learning process. The best hyperparameters after grid search are listed in Table 4.5.

Hyperparameters	**Value**
Data Augmentation Probability	0.5
Dropout Rate	0.5
Batch Size	24
Optimiser	Adam
Learning Rate	0.0001
Margin, m	0.1
Logits Scale, s	2.5
Epochs	180
Focal Loss, γ	2.0

Table 4.5: Value of hyperparameters after extensive grid-search.

5. Evaluation

This chapter attempts to present the evaluation metrics results as well as their corresponding analyses. Ablations studies of the proposed model and baseline model are reported and compared.

5.1 Testing Results

5.1.1 Single Head Setting

	AUC	Accuracy	Precision	Recall	F1
T2	0.69(±0.02)	0.63(±0.03)	0.64(±0.02)	0.63(±0.03)	0.61(±0.03)
DWI	0.77(±0.03)	0.71(±0.02)	0.71(±0.02)	0.71(±0.02)	0.71(±0.02)
ADC	0.71(±0.03)	0.64(±0.02)	0.65(±0.01)	0.64(±0.02)	0.63(±0.03)
Late Fusion	0.78(±0.01)	0.71(±0.01)	0.73(±0.01)	0.71(±0.01)	0.71(±0.01)
T2 + DWI + ADC	0.77(±0.01)	0.71(±0.02)	0.72(±0.02)	0.71(±0.02)	0.70(±0.03)

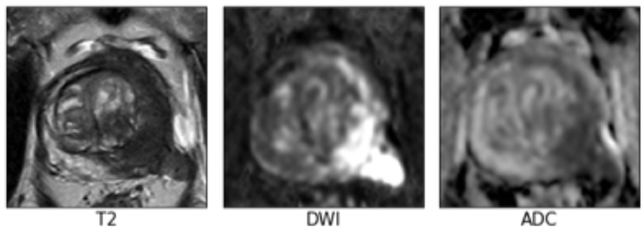
Table 5.1: Single dense head classification metrics testing data set results. Mean value (standard deviation) of the cross fold validations are shown.

Table 5.2 and Table 5.1 show the mean values and standard deviations of classification metrics for single head setting on testing data to classify patient malignancy. In both setups, inputs with DWI, late fusion and 3 image sequences as channel (T2 + DWI + ADC) yield superior performances as compared to only T2 or ADC image sequence as input. We hypothesise that due to the fact that DWI image sequences are able to display better central zone and transitional zone tumours (Lee et al. [2018]), input data affiliated with DWI contributed to the detection of cancerous tissues in the prostate. A schematic plots of T2, DWI, ADC image sequences between prostate cancer patient and healthy patient are shown in Figure 5.1. It is apparent that for prostate cancer patient (see Figure 5.1 (a)), the lesion region for DWI image shows high pixel intensity, as opposed to healthy patient (see Figure 5.1 (b)), where the

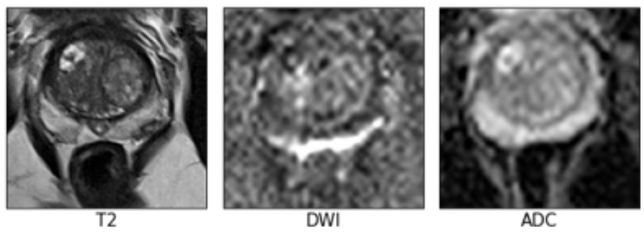
	AUC	Accuracy	Precision	Recall	F1
T2	0.70(±0.02)	0.63(±0.04)	0.65(±0.02)	0.63(±0.04)	0.62(±0.04)
DWI	0.77(±0.02)	0.70(±0.02)	0.71(±0.03)	0.70(±0.02)	0.70(±0.02)
ADC	0.73(±0.01)	0.66(±0.02)	0.66(±0.01)	0.66(±0.02)	0.65(±0.04)
Late Fusion	0.78(±0.03)	0.70(±0.02)	0.71(±0.02)	0.70(±0.02)	0.71(±0.02)
T2 + DWI + ADC	0.79(±0.01)	0.73(±0.01)	0.73(±0.02)	0.73(±0.01)	0.72(±0.01)

Table 5.2: Single ArcFace head classification metrics testing data set results. Mean value (standard deviation) of the cross fold validations are shown.

pixels intensity appear to be noisy and clustering of high pixel intensity region is not significant. We hypothesised that this discriminative feature derived from DWI images enables better learning for the model in contrast to T2 and ADC image sequences.



(a) Prostate Cancer Patient



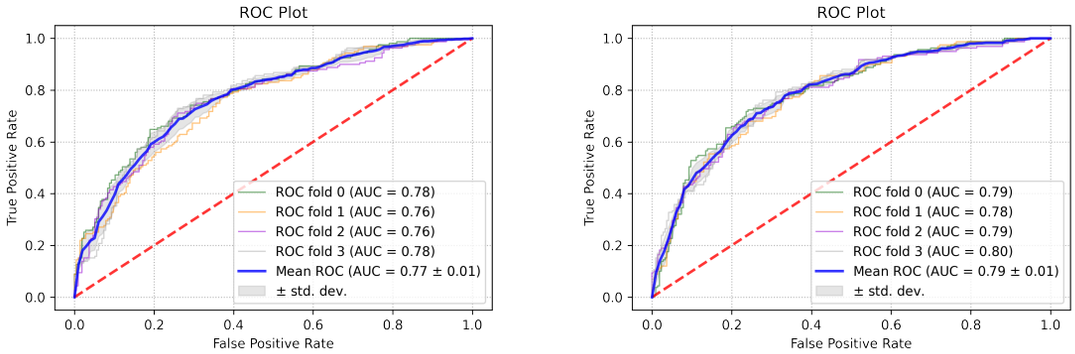
(b) Healthy Patient

Figure 5.1: Exemplary comparison plots of T2, DWI, ADC image sequences for (a) prostate cancer patient (id: 26781) and (b) healthy patient (id: 21686) retrieved from the mid slice of the patients' MRI.

By comparing the the classification metrics results on single head models from the tables, it is observed that input data with 3 image sequences as channel input (T2+DWI+ADC) for single ArcFace head yields the best results with AUC score of 0.79(±0.01), accuracy of 0.73(±0.01), precision score of 0.73(±0.02), recall score of 0.73(±0.01) and F1-score of 0.72(±0.01). Figure 5.2 shows a comparison of ROC plots for single dense head with respect to single ArcFace head. On the other hand, single dense head provides slightly better performances in DWI and late fusion inputs as compared to single ArcFace head.

For late fusion and T2+DWI+ADC settings, in both cases, the results are close to each other and in some cases, the T2+DWI+ADC setting yielded superior performance. We speculate the lack of performance from late fusion setting is due to the fact

that the concatenated embedding fails to provide distinctive features during the transfer learning. This is because the concatenated embedding depends on the feature extraction output from T2, DWI and ADC respectively, where under-performance from one of the feature extractor outputs would hamper the learning of the late fusion model. t-SNE projections of embedding plots for T2, DWI, ADC and late fusion settings are illustration in Appendix A.4. On the other hand, T2+DWI+ADC setting enables richer input information, where the model is able to learn features of the MRI image sequences at an early stage which help to boost the performance of model during the learning phase.

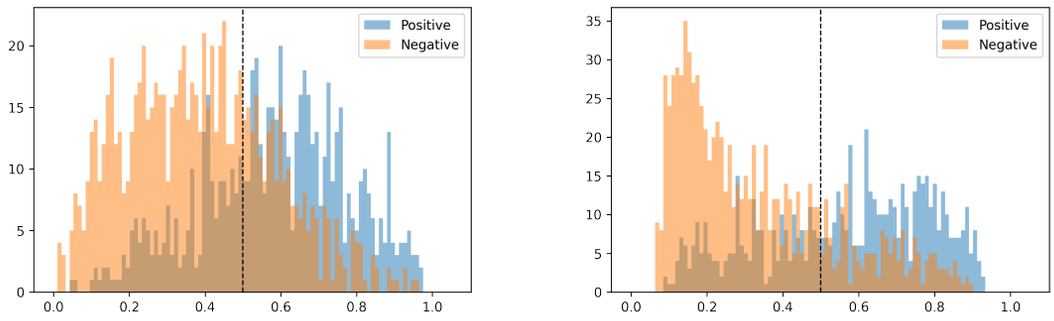


(a) Single Dense Head (T2+DWI+ADC) **(b)** Single ArcFace Head (T2+DWI+ADC)

Figure 5.2: Cross validation ROC plots for (a) single dense head and (b) single ArcFace head with AUC scores. Red dashed line represents the baseline for random classification, which is AUC of 0.50. Thick blue line represents the mean ROC plot of the validation folds. Grey shades depict the validation fold ROC standard deviations.

Figure 5.3 depicts the probability distributions of malignancy classification with T2+ DWI+ADC inputs for single dense and ArcFace heads. For single dense head, Figure 5.3 (a) shows lower confidence level for the assignment of both positive and negative labels where the probability distributions are observed to centralised in between 0.4 to 0.6 probability intervals. On the contrary, single ArcFace head shows higher confidence level for both positive and negative labels assignment, where the probability distributions are observed to taper off along the end tail of the distributions respectively. This correlates to the objective of deep metric learning where feature representations are more discriminant with similar labels by possessing higher probability confidence level. However, for ArcFace head, the maximum probability value is approximately 0.93, whereas the minimum probability value is approximately 0.07. This is due to the scaling effect of logits scale. However, high logits scale would yield higher probability confidence level, at the same time, might lead to higher mis-classification.

Table 5.3 and 5.4 list the testing results of retrieval metrics for single dense and ArcFace head respectively. It is apparent that for almost all cases ArcFace loss display superior performance in retrieval task. Other than R@1 metrics for single dense head with DWI as input having the value of 0.63($\pm$0.03), where single ArcFace head fall short of 0.03 value. The best retrieval metrics results could be observed for T2+DWI+ADC as channel input for single ArcFace head with R@1 of 0.65($\pm$0.02), R@10_Precision of 0.64($\pm$0.02) and MAP@10 of 0.51($\pm$0.03), as opposed to single



(a) Single Dense Head (T2+DWI+ADC) **(b)** Single ArcFace Head (T2+DWI+ADC)

Figure 5.3: Cross validation probability distribution for (a) single dense head and (b) single ArcFace head with AUC scores. Black dashed line refers to random classification with probability of 0.50 which is the threshold for true positives and true negatives classification.

	R@1	**R@10_Precision**	**MAP@10**
T2	0.57(±0.03)	0.55(±0.02)	0.39(±0.02)
DWI	0.63(±0.03)	0.60(±0.01)	0.45(±0.02)
ADC	0.56(±0.01)	0.56(±0.00)	0.40(±0.01)
Late Fusion	0.64(±0.03)	0.60(±0.01)	0.47(±0.02)
T2 + DWI + ADC	0.57(±0.02)	0.58(±0.02)	0.43(±0.02)

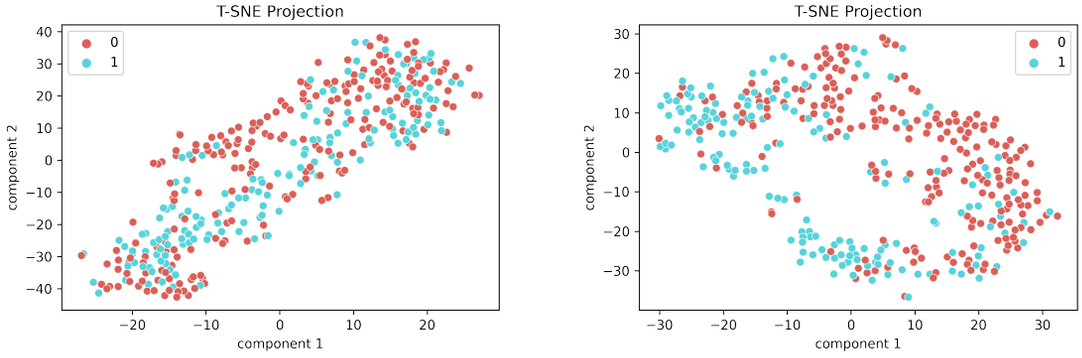
Table 5.3: Single dense head retrieval metrics testing data set results. Mean value (standard deviation) of the cross fold validations are shown.

	R@1	**R@10_Precision**	**MAP@10**
T2	0.57(±0.03)	0.56(±0.00)	0.41(±0.01)
DWI	0.60(±0.03)	0.62(±0.01)	0.47(±0.01)
ADC	0.59(±0.01)	0.58(±0.01)	0.43(±0.01)
Late Fusion	0.65(±0.04)	0.64(±0.02)	0.51(±0.03)
T2 + DWI + ADC	0.65(±0.02)	0.64(±0.02)	0.51(±0.03)

Table 5.4: Single ArcFace head retrieval metrics testing data set results. Mean value (standard deviation) of the cross fold validations are shown.

dense head with R@1 of 0.57(±0.02), R@10_Precision of 0.58(±0.02) and MAP@10 of 0.43(±0.02). The results are consistent with this thesis's hypothesis where deep metric learning, particularly ArcFace loss is able to promote class similarities while keeping dissimilar classes far from each other. The discrepancy of class similarity could be observed by mapping high dimensional embedding to low dimension visualisation with t-SNE (van der Maaten and Hinton [2008]) as illustrated in Figure 5.4 where for single ArcFace head, labels of the same class are clustered around the premises.

As opposed to single dense head in Figure 5.4 (a), separability between classes of different labels is not apparent.



(a) Single Dense Head (T2+DWI+ADC) **(b)** Single ArcFace Head (T2+DWI+ADC)

Figure 5.4: t-SNE projections of feature embedding from dimension size of 128 for (a) single dense head and (b) single ArcFace head for T2+DWI+ADC as channel input, fold 1 with perplexity of 20.

5.1.2 Multi-head Setting

	AUC	Accuracy	Precision	Recall	F1
T2	0.73(±0.02)	0.65(±0.02)	0.67(±0.02)	0.65(±0.02)	0.63(±0.03)
DWI	0.77(±0.01)	0.70(±0.02)	0.71(±0.01)	0.70(±0.02)	0.70(±0.02)
ADC	0.72(±0.02)	0.67(±0.02)	0.67(±0.03)	0.67(±0.02)	0.66(±0.01)
Late Fusion	0.77(±0.03)	0.70(±0.02)	0.71(±0.01)	0.70(±0.02)	0.69(±0.03)
T2 + DWI + ADC	0.79(±0.01)	0.71(±0.04)	0.71(±0.03)	0.71(±0.04)	0.71(±0.04)

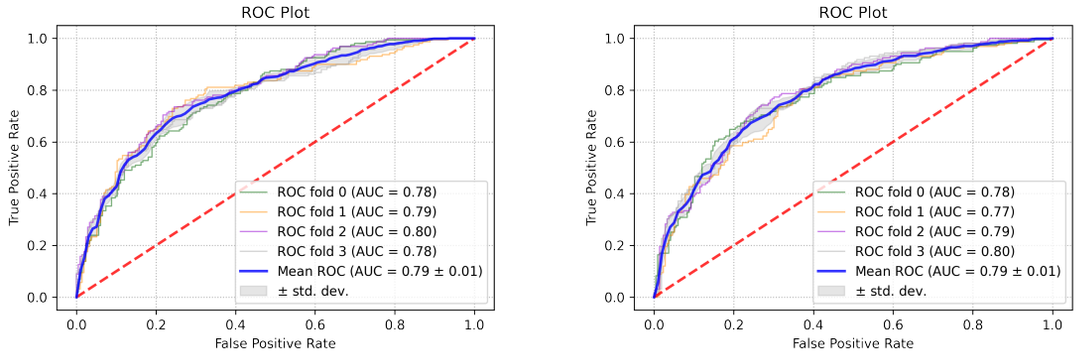
Table 5.5: Multi dense head classification metrics testing data set results. Mean value (standard deviation) of the cross fold validations are shown.

	AUC	Accuracy	Precision	Recall	F1
T2	0.72(±0.01)	0.67(±0.01)	0.66(±0.02)	0.67(±0.01)	0.66(±0.01)
DWI	0.77(±0.01)	0.70(±0.00)	0.70(±0.01)	0.70(±0.00)	0.70(±0.01)
ADC	0.71(±0.02)	0.65(±0.02)	0.65(±0.02)	0.65(±0.02)	0.65(±0.02)
Late Fusion	0.78(±0.01)	0.70(±0.01)	0.71(±0.01)	0.70(±0.01)	0.71(±0.01)
T2 + DWI + ADC	0.79(±0.01)	0.71(±0.02)	0.72(±0.03)	0.71(±0.03)	0.71(±0.02)

Table 5.6: Multi ArcFace head classification metrics testing data set results. Mean value (standard deviation) of the cross fold validations are shown.

Table 5.6 and Table 5.5 show the mean values and standard deviations of classification metrics for multi head setting on testing data. While the performances for classification metrics on both multi dense and ArcFace head are close, the best results for both

settings are observed to be T2+DWI+ADC as channel input. In this case, multi ArcFace head is slightly superior on the topping the metric scores on T2+DWI+ADC as input with AUC score of 0.79(±0.01), accuracy of 0.71(±0.02), precision score of 0.72(±0.03), recall score of 0.71(±0.03) and F1-score of 0.71(±0.02). As opposed to multi dense head with AUC score of 0.79(±0.01), accuracy of 0.71(±0.04), precision score of 0.71(±0.03), recall score of 0.71(±0.04) and F1-score of 0.71(±0.04), that fall short on 0.01 on precision score as compared to multi ArcFace head. Figure 5.5 shows a comparison of ROC plots for multi dense head with respect to multi ArcFace head for T2+DWI+ADC as channel input.

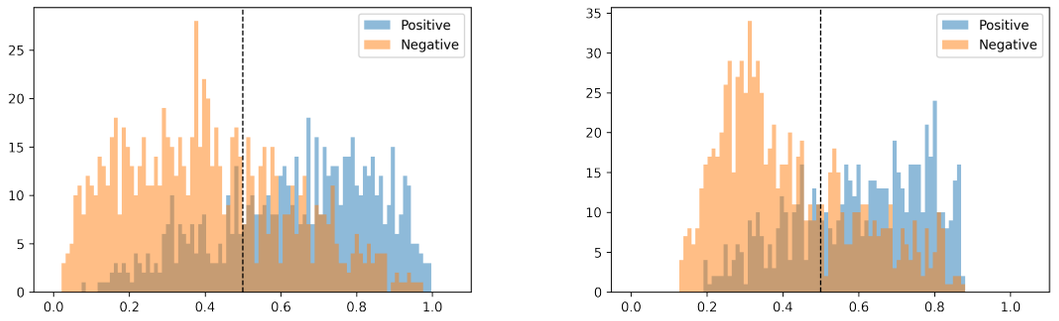


(a) Multi Dense Head (T2+DWI+ADC) **(b)** Multi ArcFace Head (T2+DWI+ADC)

Figure 5.5: Cross validation ROC plots for (a) multi dense head and (b) multi ArcFace head with AUC scores. Red dashed line represents the baseline for random classification, which is AUC of 0.50. Thick blue line represents the mean ROC plot of the validation folds. Grey shades depict the validation fold ROC standard deviations.

Figure 5.6 depicts the probability distributions of malignancy classification with T2+DWI+ADC inputs for multi dense and ArcFace heads. Unlike single dense head setting, multi dense head shows less centralised probability distribution along the 0.4 and 0.6 probability intervals. On the other hand, multi ArcFace head shows less tail distribution as compared to single ArcFace head setting probability distribution. The maximum probability value for multi ArcFace head is approximately 0.88 and the minimum proability value for multi ArcFace head is approximately 0.13. This signifies that multi ArcFace head actually hurt the performance of malignant classification where multiple heads hampered the logits from generating high confidence probability scores. This might be the reason that multi ArcFace head performed worse than single ArcFace head setting due to multiple constraints from other heads.

Table 5.7 and Table 5.8 list the testing results of retrieval metrics for multi dense and ArcFace head respectively. In all cases, multi ArcFace head shows superior performance over multi dense head for retrieval tasks. Similar to single head setting, multi ArcFace head validates the objective of deep metric learning for embedding optimisation to attract objects of similar class and repel objects of different class. The best results for multi ArcFace head retrieval metric scores are observed with T2+DWI+ADC as channel input, with R@1 of 0.65(±0.01), R@10_Precision of 0.65(±0.01) and MAP@10 of 0.55(±0.04), as opposed to multi dense head with R@1 of 0.65(±0.02), R@10_Precision of 0.62(±0.02) and MAP@10 of 0.50(±0.01). Here, the discrepancy of class similarity could be observed by mapping high dimensional



(a) Multi Dense Head (T2+DWI+ADC) **(b)** Multi ArcFace Head (T2+DWI+ADC)

Figure 5.6: Cross validation probability distribution for (a)multi dense head and (b) multi ArcFace head with AUC scores. Black dashed line refers to random classification with probability of 0.50 which is the threshold for true positives and true negatives classification.

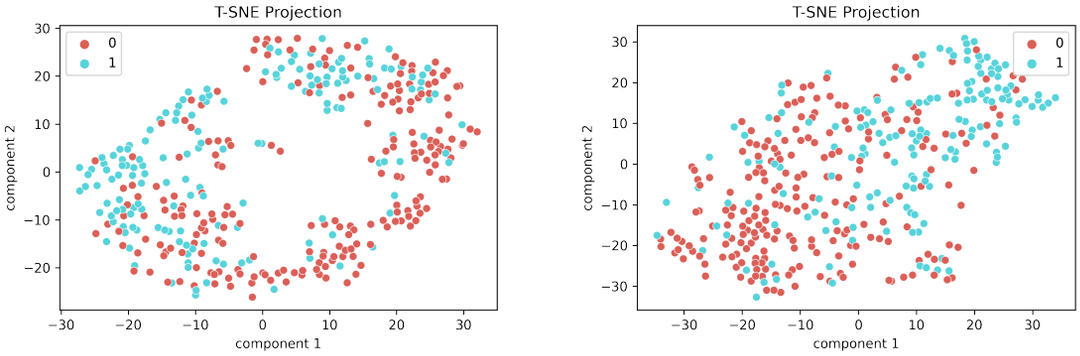
	R@1	R@10_Precision	MAP@10
T2	0.60(±0.03)	0.59(±0.02)	0.44(±0.02)
DWI	0.64(±0.04)	0.62(±0.01)	0.48(±0.03)
ADC	0.58(±0.02)	0.57(±0.02)	0.43(±0.02)
Late Fusion	0.64(±0.03)	0.63(±0.01)	0.51(±0.02)
T2 + DWI + ADC	0.65(±0.02)	0.62(±0.02)	0.50(±0.01)

Table 5.7: Multi dense head retrieval metrics testing data set results. Mean value (standard deviation) of the cross fold validations are shown.

	R@1	R@10_Precision	MAP@10
T2	0.60(±0.00)	0.59(±0.01)	0.45(±0.01)
DWI	0.64(±0.02)	0.62(±0.01)	0.50(±0.02)
ADC	0.61(±0.01)	0.59(±0.01)	0.44(±0.01)
Late Fusion	0.65(±0.01)	0.65(±0.01)	0.54(±0.01)
T2 + DWI + ADC	0.65(±0.01)	0.65(±0.01)	0.55(±0.04)

Table 5.8: Multi ArcFace head retrieval metrics testing data set results. Mean value (standard deviation) of the cross fold validations are shown.

embedding to low dimension visualisation with t-SNE (van der Maaten and Hinton [2008]) illustrated in Figure 5.7 where for multi ArcFace head, labels of the same class formed discriminant clusters. As opposed to multi dense head in Figure 5.7 (a), albeit better class separability as compared to single dense head setting, discrepancies between classes of different labels remain to be less apparent.



(a) Multi Dense Head (T2+DWI+ADC) **(b)** Multi ArcFace Head (T2+DWI+ADC)

Figure 5.7: t-SNE projections of feature embedding from dimension size of 128 for (a) multi dense head and (b) multi ArcFace head for T2+DWI+ADC as channel input, fold 1 with perplexity of 20.

5.1.3 Multi-head Outputs

In section 4.4.5, we discussed about the need for multi-head setting in order to prevent explosive combinations of classes, such that each head of the multi-head is capable of learning respective classification tasks. Instead of manifesting several latent spaces for different objectives, multi-head setting enable the navigation of one latent space for optimal learning with different classification task being assigned to one head. In this thesis work, our multi-head setting is bias on optimising an embedding for malignancy head where we aim to draw a decision boundary for biopsy decisions. On the other hand, single head setting only derive optimal learning of a latent space from a single objective. Figure 5.10 depicts visualisation plots about the distribution of the labels from prostatitis, tumour, Gleason grade group (GGG), clinical significant GGG (CSGGG) and risk head in the latent space projected from t-SNE method.

	AUC	Accuracy	Precision	Recall	F1
Prostatitis	0.65(±0.03)	0.63(±0.11)	0.66(±0.02)	0.63(±0.11)	0.61(±0.01)
Tumour	0.80(±0.03)	0.69(±0.01)	0.70(±0.02)	0.68(±0.02)	0.64(±0.07)

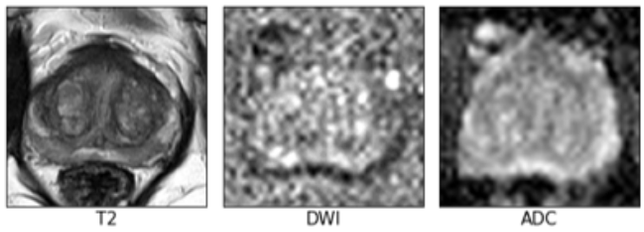
Table 5.9: Prostatitis and tumour dense head classification metrics testing data set results. Mean value (standard deviation) of the cross fold validations for T2+DWI+ADC setting are shown.

	AUC	Accuracy	Precision	Recall	F1
Prostatitis	0.66(±0.01)	0.63(±0.08)	0.70(±0.01)	0.63(±0.08)	0.64(±0.07)
Tumour	0.81(±0.02)	0.70(±0.02)	0.74(±0.02)	0.70(±0.02)	0.71(±0.01)

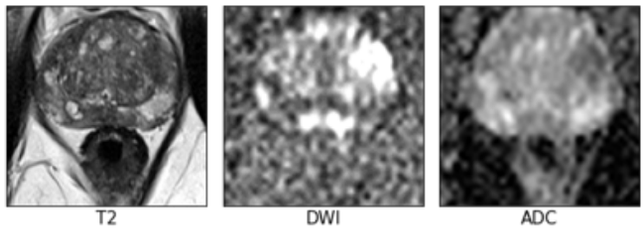
Table 5.10: Prostatitis and tumour ArcFace head classification metrics testing data set results. Mean value (standard deviation) of the cross fold validations for T2+DWI+ADC setting are shown.

It is observed in Figure 5.10 on how each head contributes to the feature embedding. Binary classes head, such as prostatitis and tumour head should provide better

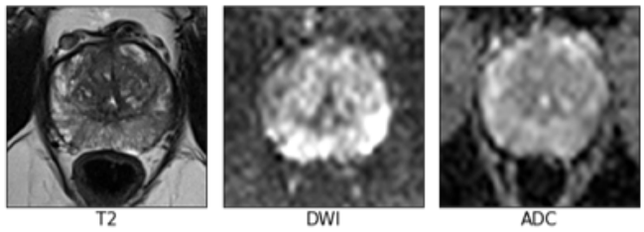
separability for the model. However, from the label distributions of the projected embedding, it is observed that tumour head provides better feature discrepancy. Unlike prostate cancer patients with lesion as feature, prostatitis patients do not possess distinct feature in the multiparametric MRI images as shown in Figure 5.8. Paticularly when a patient has both prostate cancer and prostatitis disease, the lesion region is observed to empower apparent feature. It is verified in Table 5.9 and 5.10 that tumour head is able to contribute significantly to the latent space. However, in both cases, ArcFace head provides superior performance in all classification metrics.



(a) Patient with Only Prostatitis Disease



(b) Patient with Only Prostate Cancer



(c) Patient with Prostate Cancer and Prostatitis Disease

Figure 5.8: Exemplary comparison plots of T2, DWI, ADC image sequences for (a) prostatitis patient (id: 13925), (b) prostate cancer patient (id: 17801) and (c) prostatitis and prostate cancer patient (id: 22792) retrieved from the mid slice of the patients' MRI.

On the other hand, GGG, CSGGG and risk head rely on stratification of Gleason grade which depends on Gleason patterns (see section 3.1.4.3). Gleason patterns are determined by pathologist from the prostate tissue specimen derived from whole slide images of prostate histopathology images (Linkon et al. [2021]). In the work by Lucas et al. [2019], Nagpal et al. [2019] and Pellicer-Valero et al. [2021], prostate histopathology images (see Figure 5.9) are used as input to determine the GGG of the patient. Therefore, we speculated that multiparametric MRI images in our data set do not provide ample information at a tissue specimen level for determination of Gleason pattern as opposed to histopathology images. This might be the reason why classification tasks on GGG, CSGGG and risk head do not perform well as compared to the other head. Furthermore, the lack of data set followed by imbalanced class

distribution (as shown in Figure 4.1) further jeopardised the performance of GGG, CSGGG and risk head. Concurrently, we speculated that the under-performance of these heads hinder the performance of multi-head setting as a whole as compared to single head setting.

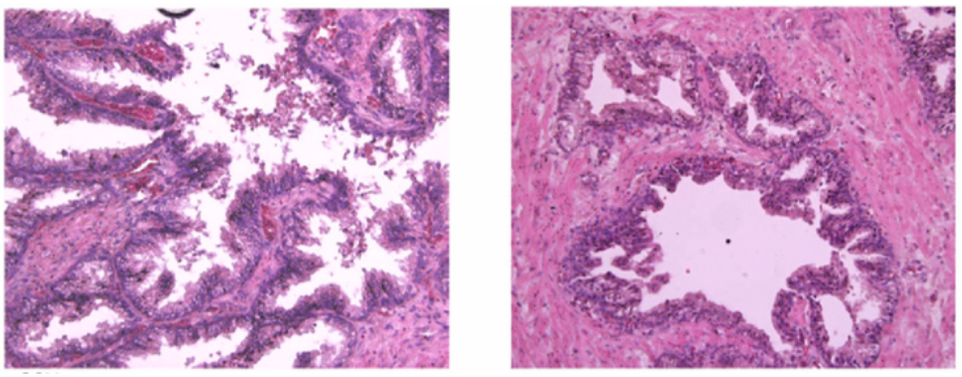


Figure 5.9: Prostate histopathology image patch retrieved from whole slice with (left) prostate cancer tissue (right) normal prostate tissue. Image source by Linkon et al. [2021].

5.2 Comparison with Related Works

	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
ImageNet by Wang et al. [2017]	0.84	-	0.70	0.84	0.79	0.77
XmasNet by Liu et al. [2017a]	0.84	-	-	-	-	-
Single Dense*	0.77	0.71	0.69	0.72	0.64	0.76
Single ArcFace*	0.79	0.73	0.63	0.79	0.69	0.75
Multi Dense*	0.79	0.71	0.73	0.68	0.63	0.77
Multi ArcFace*	0.79	0.71	0.70	0.70	0.64	0.76

Table 5.11: Comparison of prostate cancer patient differentiation statistics with ImageNet by Wang et al. [2017], XmasNet by Liu et al. [2017a] and our models* (single and multi-head dense/ArcFace settings). The cut-off value is set as 0.5. The metrics are area under ROC (AUC), accuracy, sensitivity, specificity, positive prediction value (PPV) and negative prediction value (NPV).

In section 2.2, we presented related works about computer-aided diagnosis for prostate diseases. Out of the related works, we find relevancy in the work by Wang et al. [2017] and Liu et al. [2017a] that attempts to detect prostate cancer with deep learning techniques using prostate MRI scans. Due to time constraints, we were not able to adapt our data set on the model implemented by Wang et al. [2017] (ImageNet) and Liu et al. [2017a] (XmasNet) respectively. Therefore, we only compared the

reported results of the related work with respect to our proposed models, despite the inconsistencies of data approach. Table 5.11, listed the results of the related work in comparison with our proposed models.

It is apparent in Table 5.11 that ImageNet and Xmasnet are able to achieve superior performances in AUC scores (for both models) as well as sensitivity, specificity, PPV and NPV (in the case of ImageNet). In contrast to our proposed model, both ImageNet and XmasNet took a 2D approach for prostate cancer classification task. Table 5.12 listed the data properties of ImageNet, XmasNet and our proposed methods.

ImageNet consisted of five convolution layers and two dense layer before connecting to the output layer. Each convolution layer is followed by a max-pooling layer and ReLU layer. ImageNet was implemented in the work by Wang et al. [2017] to distinguish prostate cancer patients from patients with prostatitis or benign prostatic hyperplasia (BPH) patients. Wang et al. [2017] analysed 3D data frame by frame through the usage of more than one image for each patient with the assumption that each images are independent from each other. ImageNet was pre-trained with ImageNet data set and able to achieve an impressive AUC score of 0.84 with only T2 weighted MRI image sequence as input. While it is not displayed in the paper by Wang et al. [2017] on the MRI images of prostate cancer, prostatitis and BPH patients respectively, we speculated that the T2 weighted MRI images that was used have better resolution as compared to our data set where the lesion region of the patients were highlighted. The reason ImageNet is able to perform well might be due to the less complicated network and larger sample size, viz. 2602 samples, as compared to our sample size. Not only the number of parameters are lower in ImageNet, the fact that dealing with 2D network means that larger image patch size could be implemented which prevents the lost of information, at the same time, boost the performance of the model.

On the other hand, XmasNet is a network insipired by VGG net with four convolutional layer and two dense layer before connecting to the output layer. Every convolutional layer is followed by batch a normalisation layer and a ReLU layer. Max pooling layer is only followed after every even convolutional layer. In the work by Liu et al. [2017a], the authors utilised PROSTATEx data set, which is a public available data set for prostate cancer lesion classification challenge. The authors were able to generate four inputs with the combinations of DWI (D), ADC (A), Ktrans (K) and transverse T2 (T), viz. DAK, DAT, AKT and DKT. For each type of inputs, Liu et al. [2017a] was able to generate multiple view of image slice with patch size of 32×32, having the region of interest (ROI) surrounding the lesion center. This enable XmasNet to achieve a staggering amount of sample size, i.e. 207144 samples. We speculated that since PROSTATEx is a public data set, the MRI images were well prepared. Furthermore, XmasNet is a relatively light network, due to the ROI of interest surrounding the lesion center that enables small patch size. In addition to low number of trainable model parameters, the large amount of training sample helps to tackle the overfitting problem of the model which explained why the XmasNet is able to perform well with a 0.84 AUC score.

In a nutshell, we believe that high quality data set, large amount of training sample and model with less complexity contributed to the performance of both ImageNet

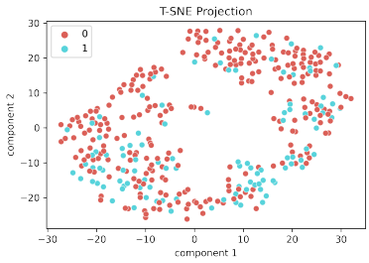
and XmasNet. While not much supportive evidence was provided in the paper by Wang et al. [2017] about such superior performance with the amount of data as compared to XmasNet, we believe that the image they possessed has distinguishable features between the prostate diseases that aided the performance of their model. Further investigation in section A.8 shows that the noise in our data set hamper the performance of our models. On the other hand, XmasNet sample size is 110 times the size of our sample size which contributes to the generalisation of the their model. Although our best models were only able to achieve a best AUC score of 0.79, we believe that our proposed model would be able to perform better with large sample size and feeding the model with higher resolution MRI images.

	Input(s)	**Patch Size**	**Sample Size**
ImageNet by Wang et al. [2017]	T2	288×288	2,602
XmasNet by Liu et al. [2017a]	Multiple	32×32	207,144
Ours*	T2+DWI+ADC	$144 \times 144 \times 24$	1,873

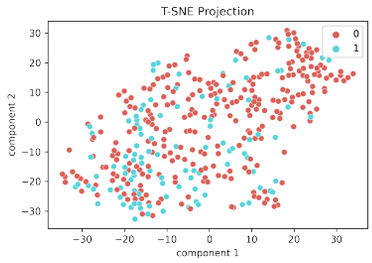
Table 5.12: Comparison of data properties with ImageNet by Wang et al. [2017], XmasNet by Liu et al. [2017a] and our models* (single and multi-head dense/ArcFace settings).

5.3 Content-based Image Retrieval (CBIR)

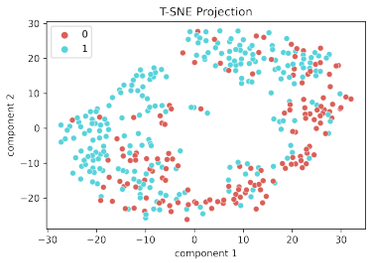
Content-based Image Retrieval (CBIR) attempts to utilise image as a query to retrieve images from the image database in replacement of keywords for more efficient retrieval process. In most work associating with CBIR (Ni et al. [2017], Zhao et al. [2021]), the retrieved image are often associated with supporting data, such as Electronic Health Record (EHR), bio-markers such as prostate PSA level, etc. However, the CBIR process is computed with 2D images. In this work, we attempt to exploit CBIR with volumetric medical images, but due to limitation of provided data, only basic information such as patient ID and class labels are provided. The potential of CBIR for 3D images could be expanded for future work with rich information for the modeling downstream tasks. In this thesis, we first map volumetric images to low dimensional embedding space, where the similarity scores between each query embedding and retrieval embedding are computed. The nearest neighbour of the query embedding are ranked according to the similarity scores in descending order. While it is not trivial to visualise volumetric data and its retrieved images, the mid slice of every patients' MRI images are chosen to be displayed for CBIR task with the assumption that the mid image slice contains the most significant image description for a patient's prostate. Exemplary images about CBIR process for single and multi head settings are shown in Figure 5.11 and Figure 5.12 respectively. It is worth noting that CBIR outcomes does not indicate the performance of retrieval metrics, but rather a showcase of possible usage of CBIR for prostate disease diagnoses in aiding doctor's decision making.



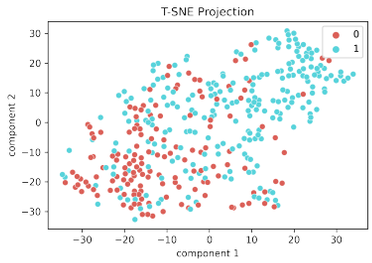
(a) Prostatitis Dense Head



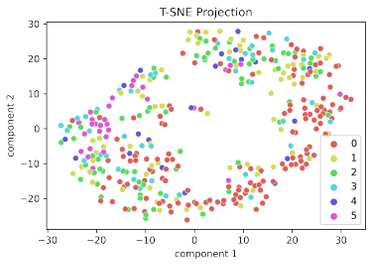
(b) Prostatitis ArcFace Head



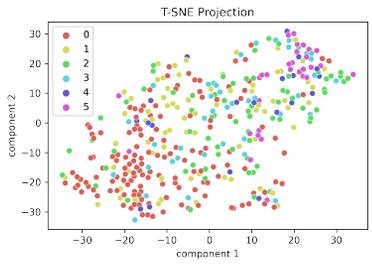
(c) Tumour Dense Head



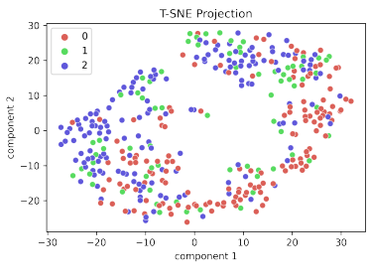
(d) Tumour ArcFace Head



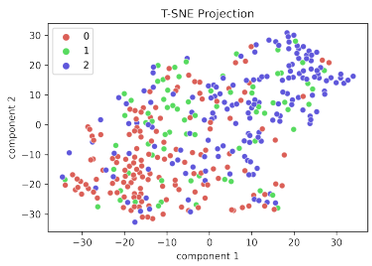
(e) GGG Dense Head



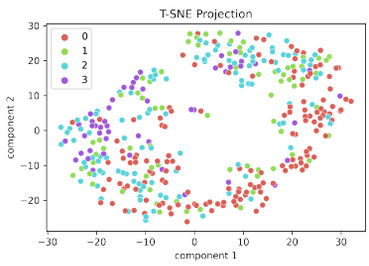
(f) GGG ArcFace Head



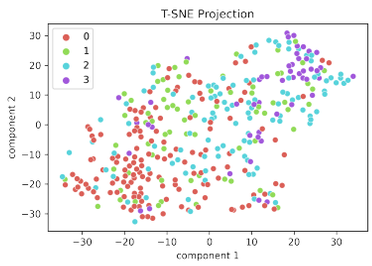
(g) CSGGG Dense Head



(h) CSGGG ArcFace Head

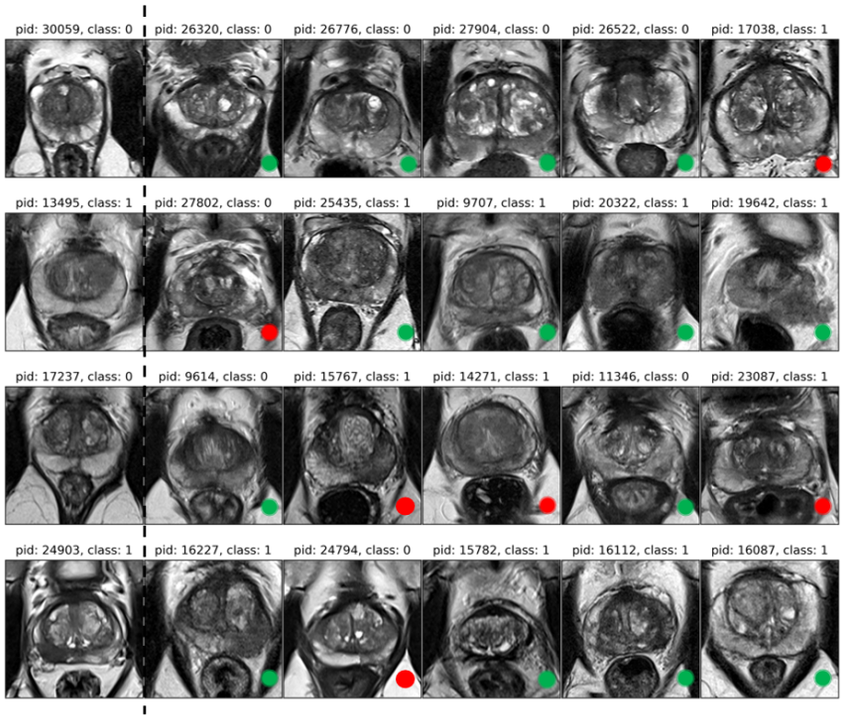


(i) Risk Dense Head

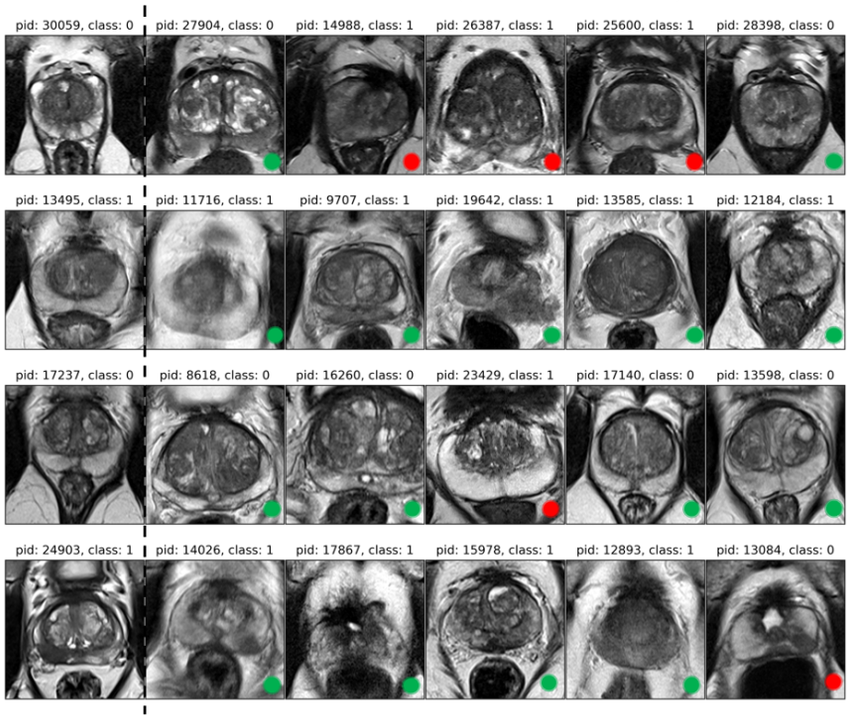


(j) Risk ArcFace Head

Figure 5.10: t-SNE projections of prostatitis, tumour, Gleason grade group (GGG), clinical significant GGG (CSGGG) and risk feature embedding for multi dense (left) and ArcFace (right) head for T2+DWI+ADC as channel input, fold 1 with perplexity of 20.

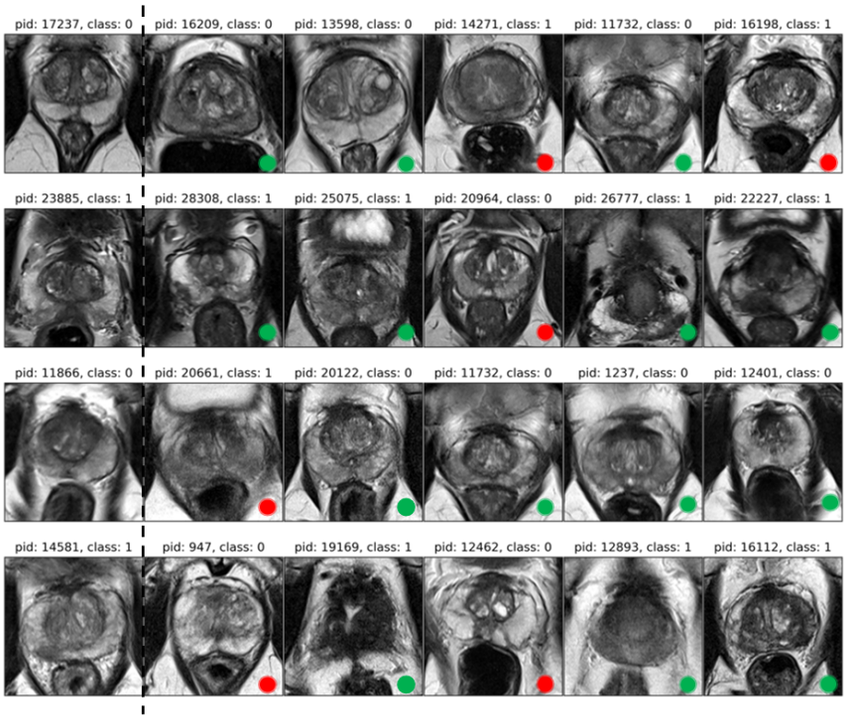


(a) Single Dense Head (T2+DWI+ADC)

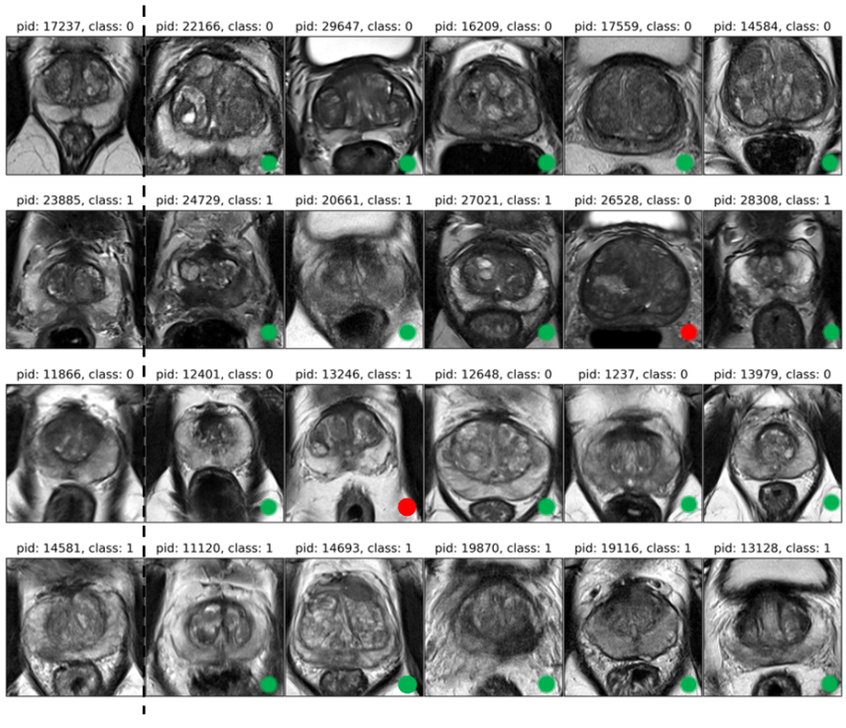


(b) Single ArcFace Head (T2+DWI+ADC)

Figure 5.11: CBIR output for (a) Single Dense Head and (b) Single ArcFace Head with T2+DWI+ADC as channel input for fold 1. Information retrieval for 4 patients are displayed. The first column are the query images. The following 5 columns are the top-5 nearest neighbours for image retrieved. The image slice for each patients belong to the middle slice. T2 image sequence are chosen as display image due to better resolution. Green dots represents correct retrieval, as opposed to red dots.



(a) Multi Dense Head (T2+DWI+ADC)



(b) Multi ArcFace Head (T2+DWI+ADC)

Figure 5.12: CBIR output for (a) Multi Dense Head and (b) Multi ArcFace Head with T2+DWI+ADC as channel input for fold 1. Information retrieval for 4 patients are displayed. The first column are the query images. The following 5 columns are the top-5 nearest neighbours for image retrieved. The image slice for each patients belong to the middle slice. T2 image sequence are chosen as display image due to better resolution. Green dots represents correct retrieval, as opposed to red dots.

6. Conclusion

This chapter presents an overview summary of the thesis and the pitfalls encountered during the thesis compilation. Potential future works will be discussed to provide future enhancements.

6.1 Discussion

In this master thesis, we presented contemporary state-of-the-art method in deep metric learning, specifically margin-based softmax loss as well as its implementations to perform medical image analysis of prostate diseases. Ongoing research in deep metric learning focuses on face recognition with 2D images. On the contrary, this thesis covered margin-based softmax loss on multiparametric-MRI (mpMRI) for computer aided diagnosis. Specifically, we presented a proposed framework of single and multi head ArcFace loss, followed by baseline setup of single and multi head softmax loss for prostate tumour malignancy classification with complimentary retrieval task. Through the literature reviews and conceptual explanation in section 2 and 3, the thesis presented investigation of the proposed method, i.e. margin-based softmax loss on medical imaging for intra-class compactness and inter-class diversity. The thesis mainly focuses on prostate tumour malignancy classification, such that the model is able to learn a decision boundary for biopsy action.

In the practical parts of the thesis in section 4 and 5, we demonstrated the implementations of our proposed model in discriminative feature learning with ArcFace loss for embedding optimisation to detect clinically significant prostate cancer. Having received mpMRI images from Altaklink, we performed analyses of the mpMRI image sequences with T2, DWI, ADC and the combination of T2+DWI+ADC as channel for the model input. In addition to these combinations, to tackle the image constraints of not being able to co-align each image sequences concurrently, we also proposed a late fusion model where the concatenation of embedding from different image sequences took place. The concatenated embedding are concatenated to form psuedo mpMRI embedding where we aim to optimise the embedding for classification task. In the implementation part, other than performing classification and retrieval task on single

head settings, we proposed multi head settings such that optimising the objectives on different head would yield better class discrepancies within the embedding space.

In the single head setting, the proposed method, ArcFace loss, achieved the best results with T2, DWI, ADC as channel input. ArcFace loss is able to provide better classification and retrieval metric scores, with AUC score of 0.79, accuracy of 0.73, R@1 of 0.65 and MAP@10 of 0.51, as opposed to softmax loss with AUC score of 0.77, accuracy of 0.71, R@1 of 0.57 and MAP@10 of 0.43. The reported metric scores verified the objective of this thesis where deep metric learning framework is able to provide better patient similarity optimisation. On the other hand, multi head ArcFace loss falls short in classification metric scores, with AUC score of 0.79 and accuracy of 0.71, as opposed to the multi head softmax loss counterpart with similar AUC score of 0.79 and accuracy of 0.71. However, multi head ArcFace loss is able to maintain better retrieval performance with R@1 of 0.65 and MAP@10 of 0.55, as compared to multi head softmax loss with R@1 of 0.65 and MAP@10 of 0.50. In a nutshell, multi head setting hurts the performance of the proposed method, ArcFace loss, where the probability distribution confidence rate is constrained in the interval of $[0.13, 0.88]$. However, multi head setting provides slight performance boost for softmax loss. In both cases, multi head setting yielded better retrieval performance for ArcFace loss and softmax loss, where it is verified that multi head settings provide better embedding optimisation. For both single and muli head settings, ArcFace loss provides discriminative feature learning to enhance higher confident rate for classification task. Furthermore, as a sub-optimal objective, this thesis also demonstrated the capability of CBIR on volumetric medical images for prostate cancer detection. Although not being able to visualise CBIR in a 3D representation as well as providing additional supportive information, the work of CBIR could be extended associating with provided information.

6.2 Challenges

Prior to margin-based softmax loss approach, we attempted contrastive approach, which is triplet loss for embedding optimisation. Other than not being able to perform classification task end-to-end for triplet loss, we suffered from the dilemma of sampling strategy for triplets mining where the model ended up with class collapse phenomenon, i.e. the model is only capable of predicting single class label. On the other hand, opting for ArcFace loss yield better results, since ArcFace loss is an extension of softmax loss through the mapping of features on a spherical embedding, at the same time, attempting to optimise the geodesic distance of similar objects with penalty margin. Despite the easier implementation and interpretation of ArcFace loss, choosing the hyperparmeter for margin and logits scale is not trivial. Too large of a logits scale value would result easier mis-classification despite having higher confidence score for class probability. Higher margin would over-penalise the objective function preventing the model from learning useful features. Last but not least, since we are dealing with private data set in this thesis, the number of data set is not sufficient enough for good model generalisation. Deeper model does not fit well in this data set where overfitting is more apparent. We have to reduce the model complexity and reside to shallow ResNet model for better generalisation.

6.3 Future Work

In section 5.2 and A.8, discussions about data noise which hinder the classification performance of our models has been pointed out. The authors of ArcFace loss, Deng et al. [2018], introduced the method with the assumption of clean training data for effective embedding optimisation. However, manually annotating and altering data set is expensive. In order to address the problem of noisy data set, an enhanced variations of state-of-the-art ArcFace loss could be implemented as future work to improve the results of the thesis's approach. Deng et al. [2020] pointed out that the original ArcFace loss is prone to label noise existing in the training data which prevents the intra-class compactness. Deng et al. [2020] proposed sub-center ArcFace, such that training sample needs to be close to one of the K sub-centers for each class instead of a class center. This approach promotes label noise robustness and leverage the need for training samples to revolve around only one center. The implementation of sub-center ArcFace is an extension of ArcFace loss, as illustrated in Figure 6.1.

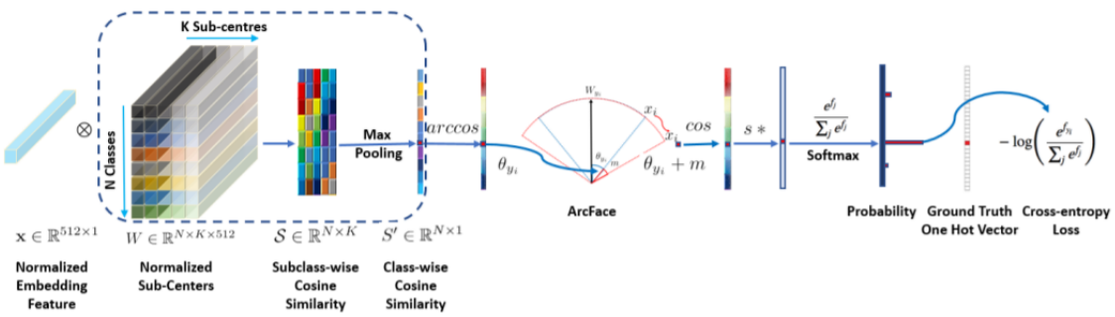


Figure 6.1: Extension of ArcFace loss (Deng et al. [2018]) with K-sub centers. Image adopted from Deng et al. [2020].

Another problem we faced in this thesis during the implementation of ArcFace loss is the difficulty of training the model at an early phase. ArcFace loss is an additive margin penalty on the angle between the features and the class center in the embedding space. While high margin value enforced better discrepancy between features of different classes, hard feature samples prevents effective model learning during the training phase. This calls for the need of sampling strategy to boost the performance of ArcFace loss. In another work by Huang et al. [2020], they incorporated mining strategies with ArcFace loss. Huang et al. [2020] proposed adaptive curriculum learning for ArcFace, namely CurricularFace, such that easy training examples are provided for the early stage of the training and progressively increasing the difficulty of training examples in the later stage of learning phase. In section 5.2, we compread the sample size of our work with respect to related works. We speculated that the lack of data samples hinder the performance of our proposed model. While it is, expensive to generate new data, work by Ko and Gu [2020] proposed an inexpensive way to generate synthetic data. The main objective of deep metric learning is to optimise the embedding space in order to achieve intra-class compactness and inter-class discrepancy. Ko and Gu [2020] proposed an augmentation method, called embedding expansion such that multiple synthetic feature points from the same class is generated. Embedding expansion by Ko and Gu [2020] is achieved by taking the linear interpolation between two feature points in

the embedding space. The benefits of this method are the ease of computation and the possibility to implement it during the training phase instead of post-processing. The aforementioned methods were not adopted because this would have exceeded the scope of this thesis.

Furthermore, the original ArcFace loss is tested extensively in face recognition tasks with large number of classes (Deng et al. [2018], Roth et al. [2020]). It is natural that softmax is used as the activation function for probability mapping in multiclass classification. However, in an addition of this work, we derived sigmoid for binary classification with ArcFace loss, shown in equation 6.1. Derivation of binary ArcFace loss is shown in Appendix A.7.

$$\begin{aligned} L = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log\left(\frac{1}{1+exp(s\,cos(\theta_{y_i}+m))}\right) + \\ (1-y_i)\cdot log\left(1-\frac{1}{1+exp(s\,cos(\theta_{y_i}+m))}\right) \end{aligned} \tag{6.1}$$

We did not implement binary ArcFace loss in this work because it was not tested extensively in any other research papers. We hypothesise that by pairing ArcFace loss with sigmoid and binary cross entropy loss, the number of model parameters in the output layer could be further reduced. Furthermore, binary ArcFace could potentially be extended to multi-label classification task. Since multi-label classification was not within the scope of this work, we were not able to demonstrate the credibility of binary ArcFace loss.

Appendix

A.1 Lower bound of Logits Scale

Given expected minimum posterior probability of class center, P_w. Assume that $cos\theta \in [-1, 1]$, we can simplified Arcface loss as shown in equation A.3 by taking the maximum values of target class, 1 else -1. C is the number of classes.

$$Arcface = \frac{exp(s\,cos(\theta_{y_i} + m))}{exp(s\,cos(\theta_{y_i} + m)) + \sum_{j \neq y_i} exp(s\,cos(\theta_j))} \tag{A.2}$$

$$\frac{exp(s)}{exp(s) + (C-1)exp(-s)} \geq P_w \tag{A.3}$$

Taking the fraction of $exp(s)$ on the left hand side of equation A.3, equation A.4 is derived.

$$\frac{1}{1 + (C+1)exp(-2s)} \geq P_w \tag{A.4}$$

Through factorisation of variables, equation A.5 is achieved.

$$\frac{1 - P_w}{P_w(C-1)} \geq exp(-2s) \tag{A.5}$$

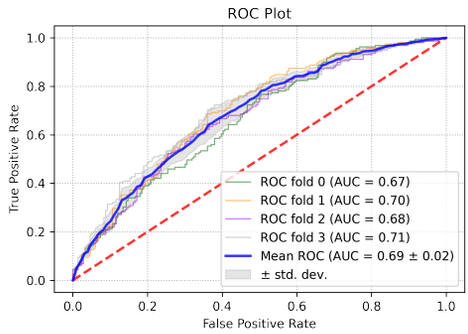
By taking natural log on both sides of equation A.5, equation A.6 is derived.

$$log\left(\frac{1 - P_w}{P_w(C-1)}\right) \geq -2S \tag{A.6}$$

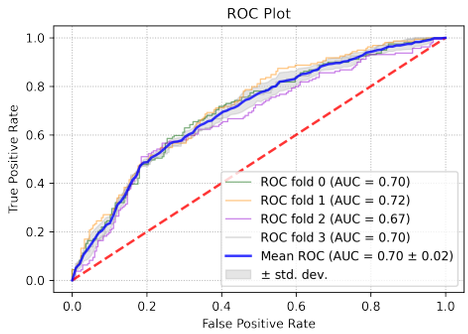
Finally, the lower bound of logits scale s is expressed as equation A.7.

$$s \geq -\frac{1}{2} log\left(\frac{1 - P_w}{P_w(C-1)}\right) \tag{A.7}$$

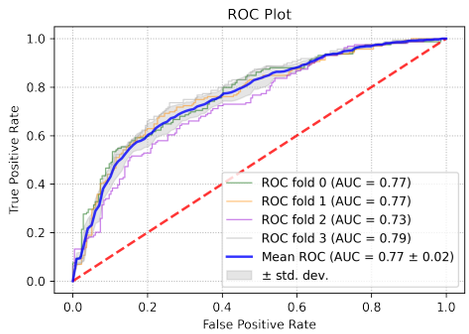
A.2 Single Head Testing Dataset AUC Plots



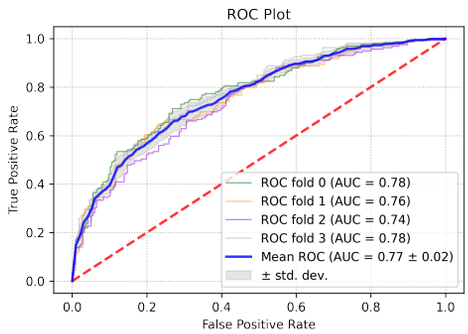
(a) Single Dense Head (T2)



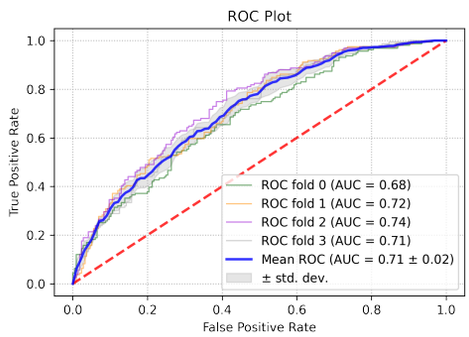
(b) Single ArcFace Head (T2)



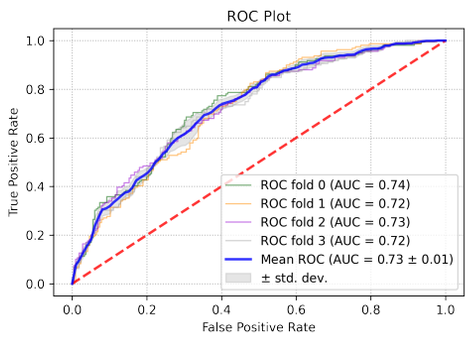
(c) Single Dense Head (DWI)



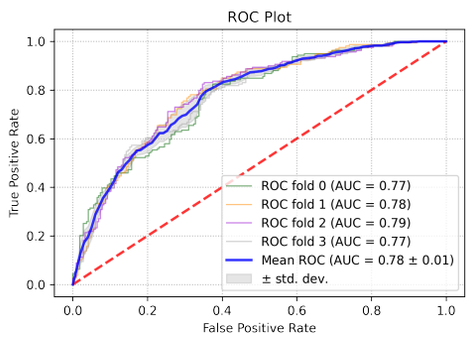
(d) Single ArcFace Head (DWI)



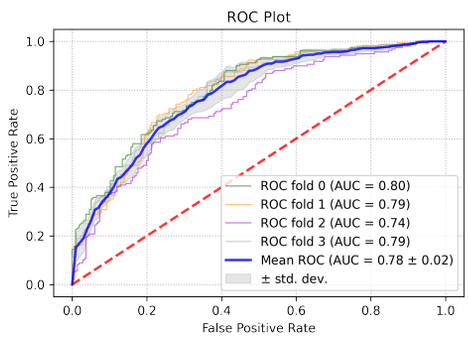
(e) Single Dense Head (ADC)



(f) Single ArcFace Head (ADC)



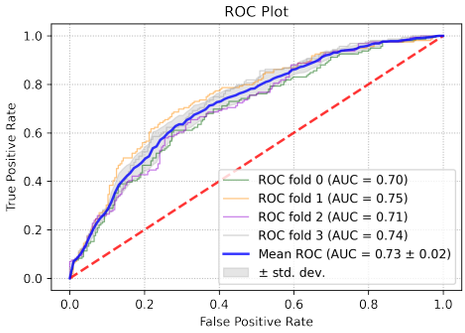
(g) Single Dense Head (Late Fusion)



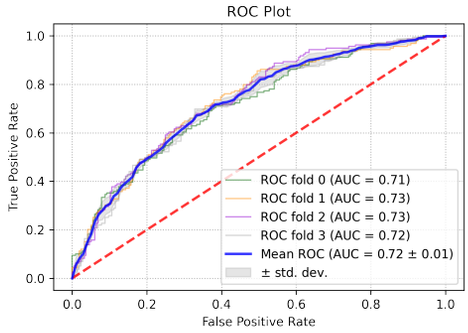
(h) Single ArcFace Head (Late Fusion)

Figure A.2: Cross validation ROC plots for single dense (left) and ArcFace (right) head with T2, DWI, ADC and Late Fusion settings.

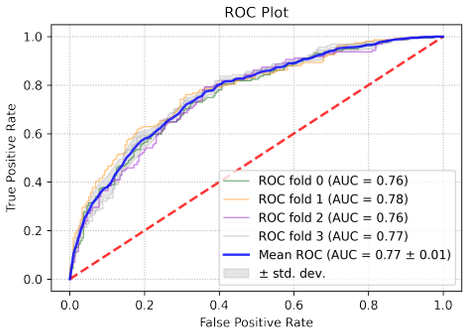
A.3 Multi Head Testing Dataset AUC Plots



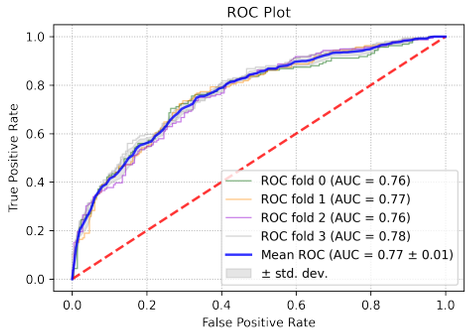
(a) Multi Dense Head (T2)



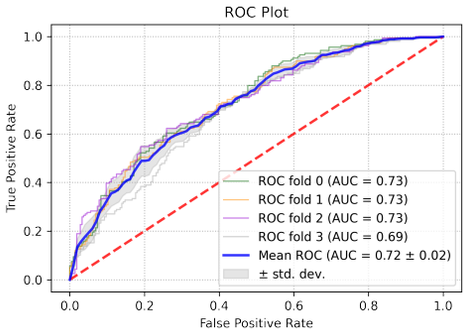
(b) Multi ArcFace Head (T2)



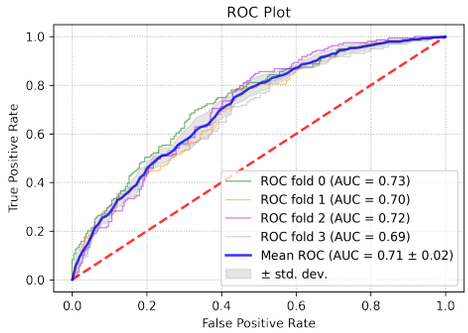
(c) Multi Dense Head (DWI)



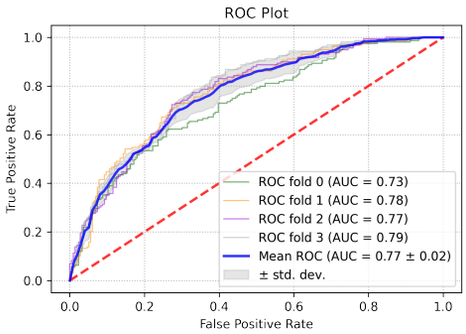
(d) Multi ArcFace Head (DWI)



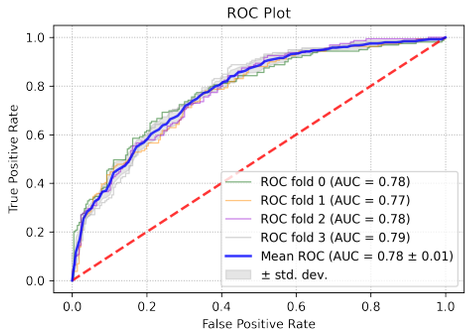
(e) Multi Dense Head (ADC)



(f) Multi ArcFace Head (ADC)



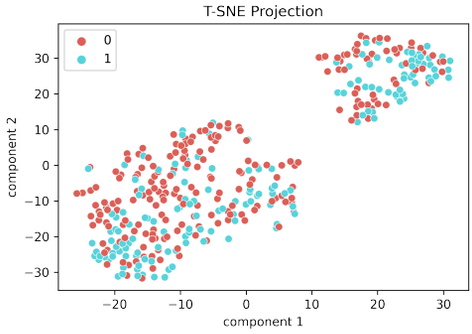
(g) Multi Dense Head (Late Fusion)



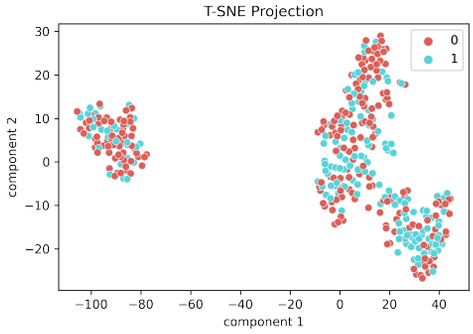
(h) Multi ArcFace Head (Late Fusion)

Figure A.3: Cross validation ROC plots for multi dense (left) and ArcFace (right) head with T2, DWI, ADC and Late Fusion settings.

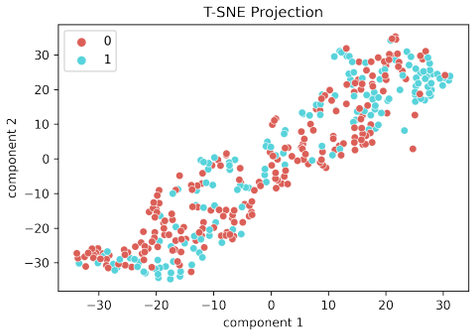
A.4 Single Head Image Sequences Embedding Plots



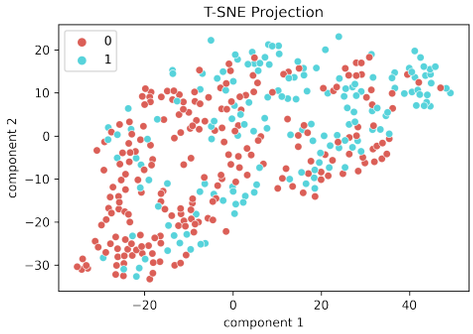
(a) Single Dense Head (T2)



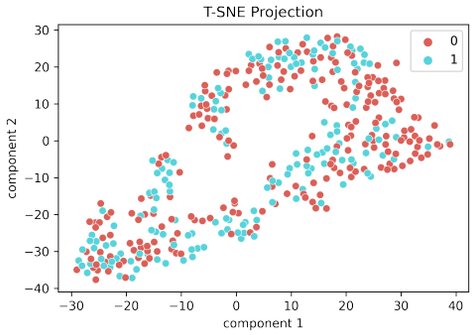
(b) Single ArcFace Head (T2)



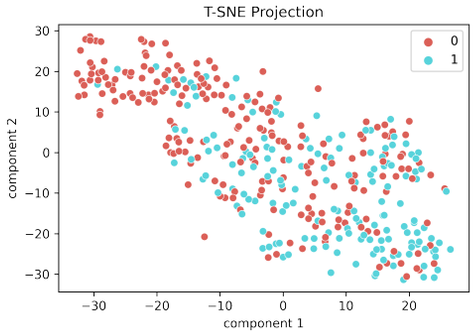
(c) Single Dense Head (DWI)



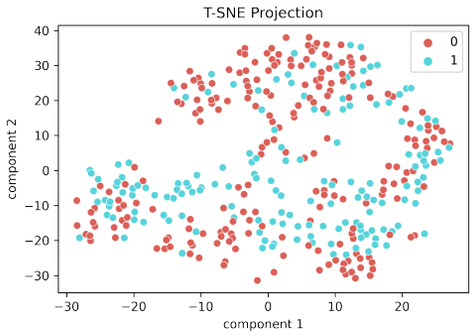
(d) Single ArcFace Head (DWI)



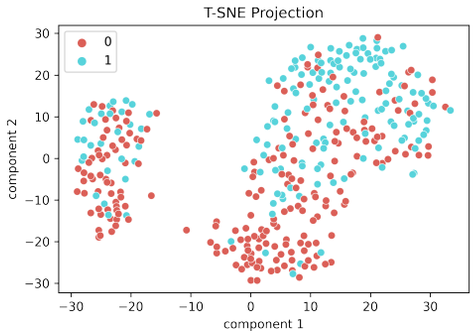
(e) Single Dense Head (ADC)



(f) Single ArcFace Head (ADC)



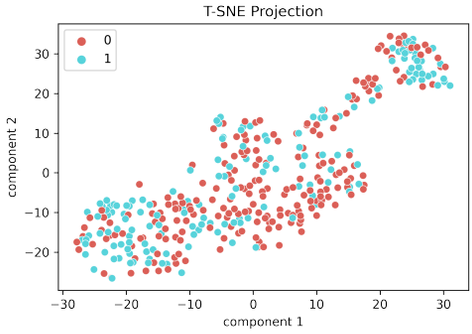
(g) Single Dense Head (Late Fusion)



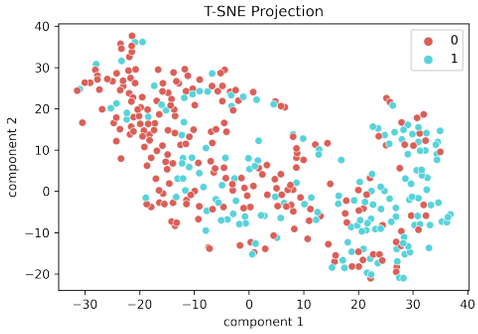
(h) Single ArcFace Head (Late Fusion)

Figure A.4: t-SNE projections of feature embedding for single dense (left) and ArcFace (right) head with T2, DWI, ADC and Late Fusion settings.

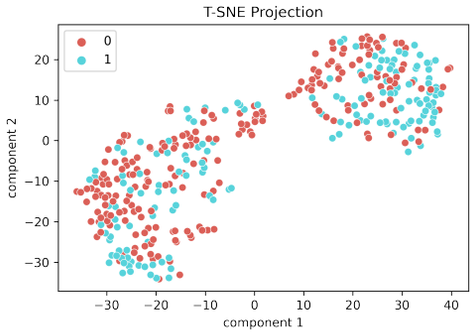
A.5 Multi Head Image Sequences Embedding Plots



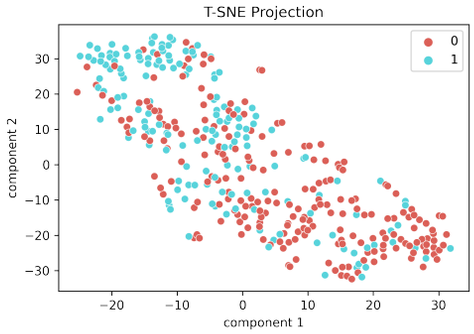
(a) Multi Dense Head (T2)



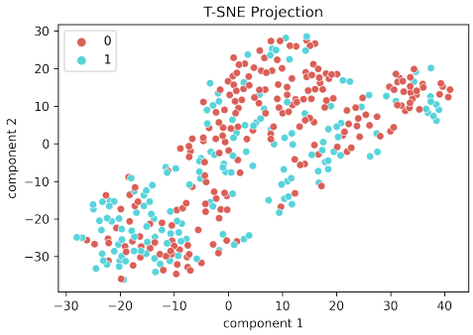
(b) Multi ArcFace Head (T2)



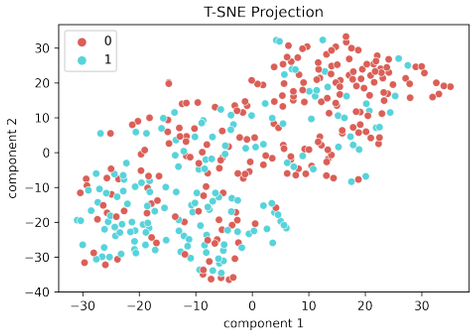
(c) Multi Dense Head (DWI)



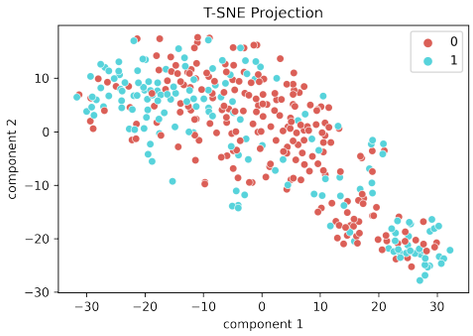
(d) Multi ArcFace Head (DWI)



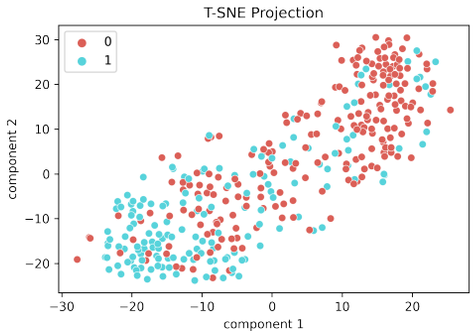
(e) Multi Dense Head (ADC)



(f) Multi ArcFace Head (ADC)



(g) Multi Dense Head (Late Fusion)



(h) Multi ArcFace Head (Late Fusion)

Figure A.5: t-SNE projections of feature embedding for multi dense (left) and ArcFace (right) head with T2, DWI, ADC and Late Fusion settings.

A.6 Combinations of Patient Granularity Classes

	Prostatitis	Malignant	GGG	Label
Healthy	0	0	0	0
Prostatitis	1	0	0	1
Prostatitis+GGG1	1	0	1	2
Prostatitis+GGG2	1	1	2	3
Prostatitis+GGG3	1	1	3	4
Prostatitis+GGG4	1	1	4	5
Prostatitis+GGG5	1	1	5	6
GGG1	0	0	1	7
GGG2	0	1	2	8
GGG3	0	1	3	9
GGG4	0	1	4	10
GGG5	0	1	5	11

Table A.1: 12 classes with combination of prostatitis diagnosis, malignant diagnosis and Gleason grade group (GGG). GGG0 refers to patient without any GGG assignments.

	Prostatitis	Malignant	Risk	Label
Healthy	0	0	No	0
Prostatitis	1	0	No	1
Prostatitis+GGG1	1	0	Low	2
Prostatitis+GGG2	1	1	Mid	3
Prostatitis+GGG3	1	1	Mid	3
Prostatitis+GGG4	1	1	High	4
Prostatitis+GGG5	1	1	High	4
GGG1	0	0	Low	5
GGG2	0	1	Mid	6
GGG3	0	1	Mid	6
GGG4	0	1	High	7
GGG5	0	1	High	7

Table A.2: 8 classes with combination of prostatitis diagnosis, malignant diagnosis and prostate Gleason grade group risk group.

	Prostatitis	Malignant	CSGGG	Label
Healthy	0	0	GGG0	0
Prostatitis	1	0	GGG0	1
Prostatitis+GGG1	1	0	GGG1	2
Prostatitis+GGG>1	1	1	GGG>1	3
GGG1	0	0	GGG1	4
GGG>1	0	1	GGG>1	5

Table A.3: 6 classes with combination of prostatitis diagnosis, malignant diagnosis and clinical significant Gleason grade group (CSGGG).

	Prostatitis	Malignant	Tumour	Label
Healthy	0	0	0	0
Prostatitis	1	0	0	1
Prostatitis+GGG1	1	0	1	2
Prostatitis+GGG>1	1	1	1	3
GGG1	0	0	1	4
GGG>1	0	1	1	5

Table A.4: 6 classes with combination of prostatitis diagnosis, malignant diagnosis and tumour diagnosis.

A.7 Binary ArcFace Loss

Sigmoid is a 2D case of softmax where we can rewrite in equation A.8.

$$\frac{e^x}{e^x + e^y} = \frac{1}{1 + e^{y-x}} \tag{A.8}$$

In binary case, we focus on positive case and neglect the negative case, where $y = 0$, equation A.8 can be rewritten as A.9.

$$\begin{aligned} \sigma(x) &= \frac{1}{1 + e^{-x}} \\ &= 1 - \sigma(y) \end{aligned} \tag{A.9}$$

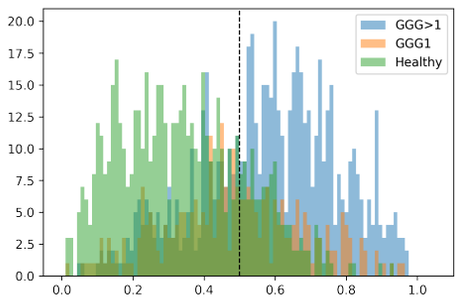
In ArcFace, target logits is expressed as, $x = s\,cos(\theta + m)$. By substituting x into binary cross entropy loss, shown in equation A.10.

$$L = -\frac{1}{N} y_i \cdot log(\sigma(x)) + (1 - y_i) \cdot log(1 - \sigma(x)) \tag{A.10}$$

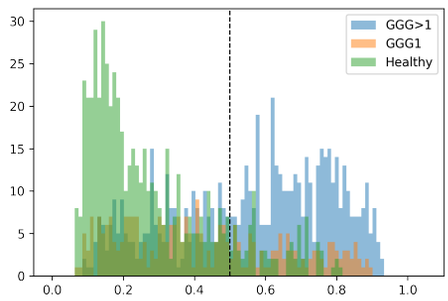
Follows, binary ArcFace loss can be expressed as A.11

$$\begin{aligned} L = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot log\left(\frac{1}{1 + exp(s\,cos(\theta_{y_i} + m))}\right) + \\ (1 - y_i) \cdot log\left(1 - \frac{1}{1 + exp(s\,cos(\theta_{y_i} + m))}\right) \end{aligned} \tag{A.11}$$

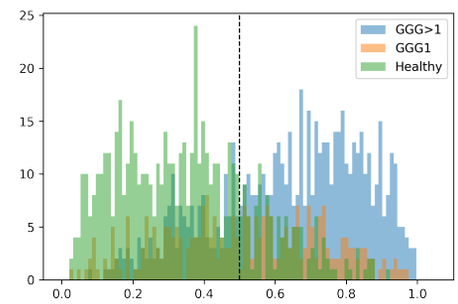
A.8 Probability Distribution of Clinical Significant Gleason Grade Group Patients



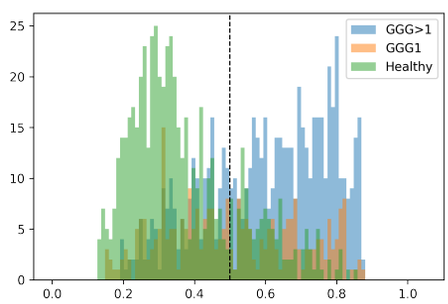
(a) Single Dense Head



(b) Single ArcFace Head



(c) Multi Dense Head



(d) Multi ArcFace Head

Figure A.6: Probability distributions of patients with clinical significant Gleason grade group (CSGGG) granularity for healthy patients, patients with GGG of 1 (GGG1) and patients with GGG larger than 1 (GGG>1) for (a) single dense head, (b) single ArcFace head, (c) multi dense head and (d) multi ArcFace head respectively.

Figure A.6 depicts the probability distributions of the classified patients from various head outputs. We further refined the class to a granularity of CSGGG where we monitor the probability distribution of healthy patients, patients with GGG of 1 (GGG1) and patients with GGG larger than 1 (GGG>1). In our work, GGG>1 patients are those whom are classified as prostate cancer patients, whereas, GGG1 patients are patients with benign tumour, which should be classified as healthy patients. From all the probability distributions, it is observed that there are a significant amount of GGG1 patients being classified as GGG>1 patients with high confidence rate. We believe these GGG1 patients MRI images are deemed as noise for our data and hampered the performance of our models.

Bibliography

Altaklinik. Prostate Cancer. https://www.altaklinik.com/prostate, accessed 2021-11-06. (cited on Page 1, 2, 10, 11, 12, and 13)

American Cancer Society. About Prostate Cancer, 2019. https://www.cancer.org/cancer/prostate-cancer/about/what-is-prostate-cancer.html, accessed 2021-11-06. (cited on Page 1, 2, 9, and 11)

Katy J.L. Bell, Chris Del Mar, Gordon Wright, James Dickinson, and Paul Glasziou. Prevalence of incidental prostate cancer: A systematic review of autopsy studies. 137(7):1749–1757, April 2015. (cited on Page 1)

Irwan Bello, William Fedus, Xianzhi Du, Ekin D. Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. *CoRR*, abs/2103.07579, 2021. URL https://arxiv.org/abs/2103.07579. (cited on Page 25)

David Bonekamp, Michael A. Jacobs, Riham El-Khouli, Dan Stoianovici, and Katarzyna J. Macura. Advancements in MR imaging of the prostate: From diagnosis to interventions. 31(3):677–703, May 2011. (cited on Page 2)

Kha Vu Chan. Deep Metric Learning: A (Long) Survey, 2021. https://hav4ik.github.io/articles/deep-metric-learning-survey, accessed 2021-11-22. (cited on Page 7 and 27)

Albert J. Chang, Karen A. Autio, Mack Roach, and Howard I. Scher. High-risk prostate cancer—classification and therapy. *Nature Reviews Clinical Oncology*, 11 (6):308–323, May 2014. doi: 10.1038/nrclinonc.2014.68. URL https://doi.org/10.1038/nrclinonc.2014.68. (cited on Page 6)

Hao Chen, Qi Dou, Lequan Yu, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. *CoRR*, abs/1608.05895, 2016. URL http://arxiv.org/abs/1608.05895. (cited on Page 5)

Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. *CoRR*, abs/1704.01719, 2017. URL http://arxiv.org/abs/1704.01719. (cited on Page 7)

Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans. *Scientific Reports*, 6(1), April 2016. doi: 10.1038/srep24454. URL https://doi.org/10.1038/srep24454. (cited on Page 5)

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. 2005. (cited on Page 3, 7, and 28)

Huseyin Cihan Demirel and John Warren Davis. Multiparametric magnetic resonance imaging: Overview of the technique, clinical applications in prostate biopsy and future directions. 44(2):93–102, February 2018. (cited on Page 12 and 13)

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009. (cited on Page 5 and 24)

Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CoRR*, abs/1801.07698, 2018. URL http://arxiv.org/abs/1801.07698. (cited on Page 3, 6, 7, 8, 30, 31, 39, 40, 67, and 68)

Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center ArcFace: Boosting face recognition by large-scale noisy web faces. In *Computer Vision – ECCV 2020*, pages 741–757. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58621-8_43. URL https://doi.org/10.1007/978-3-030-58621-8_43. (cited on Page 67)

Ketan Doshi. Batch Norm Explained Visually — How it works, and why neural networks need it, 2021. https://towardsdatascience.com/batch-norm-explained-visually-how-it-works-and-why-neural-networks-need-it-b18919692739, accessed 2021-11-20. (cited on Page 23 and 24)

Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. pages 1–1, 2021. (cited on Page 3)

Jonathan I. Epstein, Lars Egevad, Mahul B. Amin, Brett Delahunt, John R. Srigley, and Peter A. Humphrey. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma. 40(2): 244–252, February 2016. (cited on Page 13)

J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J.W.W. Coebergh, H. Comber, D. Forman, and F. Bray. Cancer incidence and mortality patterns in europe: Estimates for 40 countries in 2012. 49(6):1374–1403, April 2013. (cited on Page 1)

Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–130, January 1988. (cited on Page 20)

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *arXiv preprint arXiv:1810.12890*, 2018. (cited on Page 23)

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org. (cited on Page 15, 16, 17, 18, 19, and 21)

Google. Classification: ROC Curve and AUC, 2020. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc, accessed 2021-12-06. (cited on Page 46)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385. (cited on Page 5, 24, and 25)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016. URL http://arxiv.org/abs/1603.05027. (cited on Page 25 and 26)

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. URL http://arxiv.org/abs/1703.07737. (cited on Page 6, 7, and 29)

Vasant G Honavar. Artificial intelligence : An overview. 2014. (cited on Page 14)

Saihui Hou and Zilei Wang. Weighted channel dropout for regularization of deep convolutional neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8425–8432, July 2019. (cited on Page 23)

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL http://arxiv.org/abs/1608.06993. (cited on Page 5 and 24)

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. *CoRR*, abs/2004.00288, 2020. URL https://arxiv.org/abs/2004.00288. (cited on Page 67)

InformedHealth.org. How does the prostate work?, 2011. https://www.ncbi.nlm.nih.gov/books/NBK279291/, accessed 2021-12-06. (cited on Page 1 and 10)

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. (cited on Page 23)

Kaya and Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, August 2019. doi: 10.3390/sym11091066. URL https://doi.org/10.3390/sym11091066. (cited on Page 27)

Farhan Ullah Khan, Awais Ullah Ihsan, Hidayat Ullah Khan, Ruby Jana, Junaid Wazir, Puregmaa Khongorzul, Muhammad Waqar, and Xiaohui Zhou. Comprehensive overview of prostatitis. *Biomedicine & Pharmacotherapy*, 94: 1064–1076, October 2017. doi: 10.1016/j.biopha.2017.08.016. URL https://doi.org/10.1016/j.biopha.2017.08.016. (cited on Page 1 and 11)

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. (cited on Page 43)

ByungSoo Ko and Geonmo Gu. Embedding expansion: Augmentation in embedding space for deep metric learning. *CoRR*, abs/2003.02546, 2020. URL https://arxiv.org/abs/2003.02546. (cited on Page 67)

B. G. Vijay Kumar, Ben Harwood, Gustavo Carneiro, Ian D. Reid, and Tom Drummond. Smart mining for deep metric learning. *CoRR*, abs/1704.01285, 2017. URL http://arxiv.org/abs/1704.01285. (cited on Page 7)

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, December 1989. (cited on Page 5 and 20)

Bumshik Lee, Nagaraj Yamanakkanavar, and Jae Young Choi. Automatic segmentation of brain MRI using a novel patch-wise u-net deep architecture. *PLOS ONE*, 15(8):e0236493, August 2020. doi: 10.1371/journal.pone.0236493. URL https://doi.org/10.1371/journal.pone.0236493. (cited on Page 5)

Hakmin Lee, Sung Il Hwang, Hak Jong Lee, Seok-Soo Byun, Sang Eun Lee, and Sung Kyu Hong. Diagnostic performance of diffusion-weighted imaging for prostate cancer: Peripheral zone versus transition zone. *PLOS ONE*, 13(6):e0199636, June 2018. doi: 10.1371/journal.pone.0199636. URL https://doi.org/10.1371/journal.pone.0199636. (cited on Page 13 and 49)

Elad Levi, Tete Xiao, Xiaolong Wang, and Trevor Darrell. Reducing class collapse in metric learning with easy positive sampling. *CoRR*, abs/2006.05162, 2020. URL https://arxiv.org/abs/2006.05162. (cited on Page 7 and 29)

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. URL http://arxiv.org/abs/1708.02002. (cited on Page 44)

Ali Hasan Md. Linkon, Md. Mahir Labib, Tarik Hasan, Mozammal Hossain, and Marium-E-Jannat. Deep learning in prostate cancer diagnosis and gleason grading in histopathology images: An extensive study. *Informatics in Medicine Unlocked*, 24:100582, 2021. doi: 10.1016/j.imu.2021.100582. URL https://doi.org/10.1016/j.imu.2021.100582. (cited on Page 57 and 58)

Saifeng Liu, Huaixiu Zheng, Yesu Feng, and Wei Li. Prostate cancer diagnosis using deep learning with 3d multiparametric MRI. *CoRR*, abs/1703.04078, 2017a. URL http://arxiv.org/abs/1703.04078. (cited on Page 2, 6, 58, 59, and 60)

Siqi Liu, Daguang Xu, Shaohua Kevin Zhou, Thomas Mertelmeier, Julia Wicklein, Anna K. Jerebko, Sasa Grbic, Olivier Pauly, Weidong Cai, and Dorin Comaniciu. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. *CoRR*, abs/1711.08580, 2017b. URL http://arxiv.org/abs/1711.08580. (cited on Page 5 and 38)

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. *CoRR*, abs/1704.08063, 2017c. URL http://arxiv.org/abs/1704.08063. (cited on Page 3, 6, 7, 29, 30, and 40)

Marit Lucas, Ilaria Jansen, C. Dilara Savci-Heijink, Sybren L. Meijer, Onno J. de Boer, Ton G. van Leeuwen, Daniel M. de Bruin, and Henk A. Marquering. Deep learning for automatic gleason pattern classification for grade group determination

of prostate biopsies. *Virchows Archiv*, 475(1):77–83, May 2019. doi: 10.1007/s00428-019-02577-x. URL https://doi.org/10.1007/s00428-019-02577-x. (cited on Page 2, 6, and 57)

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Herve Delingette, Cagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, October 2015. doi: 10.1109/tmi.2014.2377694. URL https://doi.org/10.1109/tmi.2014.2377694. (cited on Page 5)

David C. Miller, Khaled S. Hafez, Andrew Stewart, James E. Montie, and John T. Wei. Prostate carcinoma presentation, diagnosis, and staging. 98(6):1169–1178, September 2003. (cited on Page 2)

MLee. Visual Guide to the Confusion Matrix, 2021. https://towardsdatascience.com/visual-guide-to-the-confusion-matrix-bb63730c8eba, accessed 2021-11-27. (cited on Page 45)

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. A metric learning reality check. *CoRR*, abs/2003.08505, 2020. URL https://arxiv.org/abs/2003.08505. (cited on Page 46)

Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L. Smith, Arash Mohtashamian, James H. Wren, Greg S. Corrado, Robert MacDonald, Lily H. Peng, Mahul B. Amin, Andrew J. Evans, Ankur R. Sangoi, Craig H. Mermel, Jason D. Hipp, and Martin C. Stumpe. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *npj Digital Medicine*, 2(1), June 2019. doi: 10.1038/s41746-019-0112-2. URL https://doi.org/10.1038/s41746-019-0112-2. (cited on Page 2, 6, and 57)

Leen Naji, Harkanwal Randhawa, Zahra Sohani, Brittany Dennis, Deanna Lautenbach, Owen Kavanagh, Monica Bawor, Laura Banfield, and Jason Profetto. Digital rectal examination for prostate cancer screening in primary care: A systematic review and meta-analysis. 16(2):149–154, March 2018. (cited on Page 2)

National Cancer Institute. Prostate Cancer Morphology and Grade. https://training.seer.cancer.gov/prostate/abstract-code-stage/morphology.html, accessed 2021-11-15. (cited on Page 13)

Jiazhi Ni, Jie Liu, Chenxin Zhang, Dan Ye, and Zhirou Ma. Fine-grained patient similarity measuring using deep metric learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, November 2017. doi: 10.1145/3132847.3133022. URL https://doi.org/10.1145/3132847.3133022. (cited on Page 3, 7, 8, and 60)

Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018. URL http://arxiv.org/abs/1804.03999. (cited on Page 5)

Yuto Onga, Shingo Fujiyama, Hayato Arai, Yusuke Chayama, Hitoshi Iyatomi, and Kenichi Oishi. Efficient feature embedding of 3d brain MRI images for content-based image retrieval with deep metric learning. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, December 2019. doi: 10.1109/bigdata47090.2019.9006364. URL https://doi.org/10.1109/bigdata47090.2019.9006364. (cited on Page 7 and 8)

Anabik Pal, Zhiyun Xue, Brian Befano, Ana Cecilia Rodriguez, L. Rodney Long, Mark Schiffman, and Sameer Antani. Deep metric learning for cervical image classification. *IEEE Access*, 9:53266–53275, 2021. doi: 10.1109/access.2021.3069346. URL https://doi.org/10.1109/access.2021.3069346. (cited on Page 3 and 8)

Amit R Patel and J Stephen Jones. Optimal biopsy strategies for the diagnosis and staging of prostate cancer. 19(3):232–237, May 2009. (cited on Page 2)

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. (cited on Page 23)

Oscar J. Pellicer-Valero, José L. Marenco Jiménez, Victor Gonzalez-Perez, Juan Luis Casanova Ramón-Borja, Isabel Martín García, María Barrios Benito, Paula Pelechano Gómez, José Rubio-Briones, María José Rupérez, and José D. Martín-Guerrero. Deep learning for fully automatic detection, segmentation, and gleason grade estimation of prostate cancer in multiparametric magnetic resonance images, 2021. (cited on Page 6 and 57)

Islam Reda, Ashraf Khalil, Mohammed Elmogy, Ahmed Abou El-Fetouh, Ahmed Shalaby, Mohamed Abou El-Ghar, Adel Elmaghraby, Mohammed Ghazal, and Ayman El-Baz. Deep learning role in early diagnosis of prostate cancer. *Technology in Cancer Research & Treatment*, 17:153303461877553, January 2018. doi: 10.1177/1533034618775530. URL https://doi.org/10.1177/1533034618775530. (cited on Page 2 and 6)

Oren Rippel, Manohar Paluri, Piotr Dollar, and Lubomir Bourdev. Metric learning with adaptive density discrimination, 2015. (cited on Page 7)

Claus G Roehrborn. Benign prostatic hyperplasia: An overview. *Rev Urol*, 2005. (cited on Page 1 and 10)

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597. (cited on Page 5)

Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Björn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. *CoRR*, abs/2002.08473, 2020. URL https://arxiv.org/abs/2002.08473. (cited on Page 29, 31, and 68)

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015. URL http://arxiv.org/abs/1503.03832. (cited on Page 3, 6, 7, 28, and 29)

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), July 2019. (cited on Page 22)

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. (cited on Page 5 and 24)

J. Smith-Palmer, C. Takizawa, and W. Valentine. Literature review of the burden of prostate cancer in germany, france, the united kingdom and canada. 19(1), March 2019. (cited on Page 1 and 2)

Anuroop Sriram, Matthew J. Muckley, Koustuv Sinha, Farah Shamout, Joelle Pineau, Krzysztof J. Geras, Lea Azour, Yindalon Aphinyanaphongs, Nafissa Yakubova, and William Moore. COVID-19 prognosis via self-supervised representation learning and multi-image prediction. *CoRR*, abs/2101.04909, 2021. URL https://arxiv.org/abs/2101.04909. (cited on Page 6)

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. (cited on Page 22 and 23)

Josefine Vilsbøll Sundgaard, James Harte, Peter Bray, Søren Laugesen, Yosuke Kamide, Chiemi Tanaka, Rasmus R. Paulsen, and Anders Nymark Christensen. Deep metric learning for otitis media classification. *Medical Image Analysis*, 71: 102034, July 2021. doi: 10.1016/j.media.2021.102034. URL https://doi.org/10.1016/j.media.2021.102034. (cited on Page 3 and 8)

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. URL http://arxiv.org/abs/1409.4842. (cited on Page 5 and 24)

Octavian Sabin Tătaru, Mihai Dorin Vartolomei, Jens J. Rassweiler, Oșan Virgil, Giuseppe Lucarelli, Francesco Porpiglia, Daniele Amparore, Matteo Manfredi, Giuseppe Carrieri, Ugo Falagario, Daniela Terracciano, Ottavio de Cobelli, Gian Maria Busetto, Francesco Del Giudice, and Matteo Ferro. Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives. 11(2):354, February 2021. (cited on Page 2)

Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015. (cited on Page 23)

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html. (cited on Page 52 and 55)

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *CoRR*, abs/1801.09414, 2018. URL http://arxiv.org/abs/1801.09414. (cited on Page 3, 6, 8, 30, and 40)

Xinggang Wang, Wei Yang, Jeffrey Weinreb, Juan Han, Qiubai Li, Xiangchuang Kong, Yongluan Yan, Zan Ke, Bo Luo, Tao Liu, and Liang Wang. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. 7(1), November 2017. (cited on Page 2, 5, 6, 58, 59, and 60)

Xun Wang, Xintong Han, Weiling Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. *CoRR*, abs/1904.06627, 2019. URL http://arxiv.org/abs/1904.06627. (cited on Page 8)

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision – ECCV 2016*, pages 499–515. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46478-7_31. URL https://doi.org/10.1007/978-3-319-46478-7_31. (cited on Page 7 and 30)

Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. *CoRR*, abs/1706.07567, 2017. URL http://arxiv.org/abs/1706.07567. (cited on Page 31)

Lian-Ming Wu, Jian-Rong Xu, Yong-Quan Ye, Qing Lu, and Jia-Ni Hu. The clinical value of diffusion-weighted imaging in combination with t2-weighted imaging in diagnosing prostate carcinoma: A systematic review and meta-analysis. 199(1): 103–110, July 2012. (cited on Page 12)

Shu-Jie Xia, Di Cui, and Qi Jiang. An overview of prostate diseases and their characteristics specific to asian men. *Asian Journal of Andrology*, 14(3):458–464, February 2012. doi: 10.1038/aja.2010.137. URL https://doi.org/10.1038/aja.2010.137. (cited on Page 1)

Sangjun Yoo, Jeong Kon Kim, and In Gab Jeong. Multiparametric magnetic resonance imaging for prostate cancer: A review and update for urologists. 56(7): 487, 2015. (cited on Page 12)

Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. *CoRR*, abs/1905.00292, 2019. URL http://arxiv.org/abs/1905.00292. (cited on Page 40)

Wentao Zhao, Wei Jiang, and Xinguo Qiu. Deep learning for COVID-19 detection based on CT images. *Scientific Reports*, 11(1), July 2021. doi: 10.1038/s41598-021-93832-2. URL https://doi.org/10.1038/s41598-021-93832-2. (cited on Page 6 and 60)

Aoxiao Zhong, Xiang Li, Dufan Wu, Hui Ren, Kyungsang Kim, Younggon Kim, Varun Buch, Nir Neumark, Bernardo Bizzo, Won Young Tak, Soo Young Park, Yu Rim Lee, Min Kyu Kang, Jung Gil Park, Byung Seok Kim, Woo Jin Chung, Ning Guo, Ittai Dayan, Mannudeep K. Kalra, and Quanzheng Li. Deep metric learning-based image retrieval system for chest radiograph and its clinical applications in COVID-19. 70:101993, May 2021. (cited on Page 6, 7, 8, and 31)

Zhou and Chellappa. Computation of optical flow using a neural network. In *IEEE International Conference on Neural Networks*. IEEE, 1988. (cited on Page 21)

Mu Zhou, Yusuke Tanimura, and Hidemoto Nakada. One-shot learning using triplet network with kNN classifier. In *Advances in Intelligent Systems and Computing*, pages 227–235. Springer International Publishing, 2020. doi: 10.1007/978-3-030-39878-1_21. URL https://doi.org/10.1007/978-3-030-39878-1_21. (cited on Page 3)

Nor Asma Mohd Zin, Rozianiwati Yusof, Saima Anwar Lashari, Aida Mustapha, Norhalina Senan, and Rosziati Ibrahim. Content-based image retrieval in medical domain: A review. *Journal of Physics: Conference Series*, 1019:012044, June 2018. doi: 10.1088/1742-6596/1019/1/012044. URL https://doi.org/10.1088/1742-6596/1019/1/012044. (cited on Page 8)

STATEMENT OF AUTHORSHIP

Thesis: Classification of Clinically Significant Prostate Cancer with Multiparametric MRI

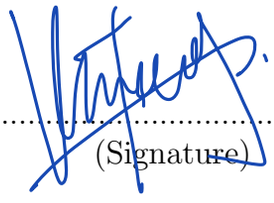
Name: Wai Po Kevin Teng **Matriculation no.:** 221219

Date of birth: 27.07.1993

I herewith assure that I wrote the present thesis independently, that the thesis has not been partially or fully submitted as graded academic work and that I have used no other means than the ones indicated. I have indicated all parts of the work in which sources are used according to their wording or to their meaning.

I am aware of the fact that violations of copyright can lead to injunctive relief and claims for damages of the author as well as a penalty by the law enforcement agency.

Magdeburg, 04.01.2022: .....



(Signature)