Classification of Small Renal Masses Based on CT Images and Machine Learning Algorithms

Master's thesis (Tech.)
University of Turku
Department of Future Technologies
2018
Heyang Sun

Supervisors:
M.Sc. Jussi Toivonen
Assco. Prof. Guohui Zhou

UNIVERSITY OF TURKU
Department of Future Technologies

SUN HEYANG: Classification of Small Renal Masses Based on CT Images and Machine Learning Algorithms
Master of Science in Technology Thesis, 58 p.
Computer Systems
May 2018

Kidney tumor is among the leading causes of tumors and deaths worldwide. In all kidney tumor cases, an increasing number of small renal masses (SRMs) with a size smaller than 4 cm have been detected and they are becoming a typical problem for radiologists and surgeons. Most SRMs are either of renal angiomyolipoma (AML) or renal cell carcinoma (RCC), the former being benign and the latter being malignant. The malignant ones can be further classified into three types, clear cell renal cell carcinoma (ccRCC), papillary renal cell carcinoma (pRCC), and chromophobe renal cell carcinoma (chRCC). Different kind of renal tumor requires varied treatment and management.

In recent years, four-phase computer tomography (CT) has become the standard approach for kidney tumor examination. In most circumstances, classic AMLs and RCCs can be classified by a radiologist reading the CT images. While fat poor angiomyolipomas (fp-AML) set barriers to this classification method due to the loss of typical diagnosis characteristics. Radiologists are also incapable of differentiating malignant tumors. For now, SRM classification is mainly performed by pathological examination, which is time and resource consuming.

Machine learning and one of its branch, deep learning, has been extended to medical image processing field. In this paper, support vector machine (SVM) and convolutional neural network (CNN) were respectively used to build models with the input of one of the last three phases of CT images and the combination of them. For the establishment of each model, at least 20% of overall patient cases were picked out randomly as independent testing subset and the rest undertook 10-fold cross validation for an objective and reliable evaluation of the models.

It turned out that SVM algorithm using a linear kernel with phase 2 (corticomedullary) images as input acquired an accuracy of 0.93 and a sensitivity of 0.97 on patient's tumor type prediction of fp-AML/RCC classification. CNN algorithm, consisting of 12 layers including 4 convolutional layers each followed by a max-pooling layer, one flatten layer, and three densely connected layers, with the help of activation functions, dropout strategy, and stochastic gradient descent (SGD) optimization method, achieved an accuracy of 0.85 on pRCC/chRCC/ccRCC categorization with phase 2 images as input. Images of corticomedullary stage were proved to be eligible for classifiers. This can be seen as a breakthrough since it is the first successful application of deep learning networks in renal tumor classification. Meanwhile, these two models were both balanced over different

classes and they together provide a comprehensive solution to SRM classification.

Given these findings, the two models can be a preliminary step for machine learning and especially deep learning algorithms to assist, improve, and finally revolutionize the conventional clinical decision making process to guide appropriate management and treatment.

Keywords: SRM; Classification; Four-phase CT images; SVM; CNN

Contents

Chapter 1

Introduction

1.1 Background of the Research

Kidney cancer is among the 10 leading causes of cancer cases and deaths in developed countries [1]. Meanwhile, in year 2009-2011, the case rate of kidney cancer for Chinese people is 66.8 per 100,000 while the mortality is 23.4 per 100,000. Both the case rate and death rate of men with kidney cancer is about twice those of women [2]. In all kidney cancer cases, an increasing number of more small and clinically localized renal masses with a size smaller than 4cm have been detected and these small renal masses (SRMs) are becoming a typical problem for radiologists and surgeons [3]. Most SRMs are either of renal angiomyolipoma (AML) type or renal cell carcinoma (RCC) type. AML are benign tumors which occur infrequently in the general population while RCC represents 2-3% of adult cancers and accounts for 90% of renal malignancies, and is regarded as the most lethal neoplasm of the urologic system [4]. An annual increase of approximately 2% has also been witnessed in worldwide incidence [5]. RCC can be further categorized into several distinct subtypes, the leading three classes being clear cell renal cell carcinoma (ccRCC), papillary renal cell carcinoma (pRCC), and chromophobe renal cell carcinoma (chRCC) [4].

Different categories of kidney cancer demand distinctive ways of postoperative management and determine deviating life spans of patients. More specifically, AML don’t

require much care after the resection surgery while RCC call for postoperative check and involvement of chemotherapy and/or radiotherapy (in most circumstances, pRCC and chRCC are in need of only chemotherapy, while ccRCC need both). In addition, the postoperative supervision of RCC all involve follow-ups. For pRCC and chRCC, occasional follow-ups are adequate but for ccRCC, more frequent follow-ups are significant. As for life span, ccRCC cannot expect a longer survival than pRCC and chRCC. This is why the diagnosis of tumor type is imperative.

Imaging technologies, especially computer tomography (CT) and magnetic resonance imaging (MRI), have become necessary approaches for renal mass detection which led to the growth of renal tumor incidence rate [6] and also plays an important role in tumor resection by providing crucial information of tumor size, shape, and location. It is through the pathological examination of the excised tissue after the surgery that the correct and convincing diagnosis can be obtained. Considering the fact that CT/MRI is an indispensable step for renal tumor treatment, and that pathological examination is time and human resources consuming, the idea of diagnosis via CT/MRI has gathered great momentum in last a few years on tumors including lung cancer, liver cancer, etc.

Machine learning algorithms are the mainstream image processing methods in recent years and has been proved effective, time, energy and human resources saving and reliable on the analysis of medical images. Deep learning, a branch of machine learning, has also grown more and more popular with researchers. The classification of small renal masses by the application of machine learning including deep learning algorithms on CT images is a topic that hasn't been covered by any other studies, and it is exactly the focus of my work. I brought two state-of-art algorithms which are SVM and CNN for the CT images analysis and they proved to be well-functioning and superior performing.

1.2 Small Renal Masses

As mentation in 1.1, small renal masses can be classified into benign tumors and malignant tumors. Malignant tumors can be subdivided into chRCC, pRCC, and ccRCC.

A more detailed description of these tumors is related in this section.

An AML must include varied amounts of three components that are dysmorphic blood vessels, smooth muscle and mature adipose tissue [7]. The classic AML have a low attenuation component resulted from a certain amount of fat content, which makes it possible to detect AML correctly with CT or MRI imaging technologies [8]. However, there are also a great number of AML, called fat poor angiomyolipomas (fp-AML), in lacking of evident fat cells, which may lead to false preoperative diagnosis by imaging technologies [8][9]. These failures of mistakenly diagnosing an AML to be an RCC can contribute to unnecessary operations. In consideration of these failures, accurate differentiation of fp-AML and RCC is of great importance to avoid dispensable surgeries.

RCC can be subdivided into six clinically and histologically different malignant tumor types, all arising from the nephron, which is the basic functional unit of the kidney. ccRCC and pRCC, originating in the proximal tubule, add up to 90% of all cases while chRCC, arising in the distal tubule, account for 5% of all cases. Other types, which are renal oncocytomas, collecting duct and medullary RCC, each represent less than 1% of all cases [10].

The risk factors which have an influence on the tumor to be benign or malignant include patient age, gender, tumor size, and smoking history [11].

1.3 Renal Masses Diagnostic Techniques

1.3.1 Current Situation of SRM Diagnostic and Treatment Methods

For renal masses diagnosis, there are both radiological and pathological approaches since different classes of renal masses relates to different pathology and clinical behavior. To understand their difference is essential not only in clinical practice but also for academic research. Radiological examination works by identifying imaging features of more fat cells and smaller size to tell AML from RCC [12]. Due to the existence of fp-AML, the radiological decision is often not dependable enough. Pathological examination is the golden standard for renal mass classification and always involves taking advantage

of histological appearance, molecular biology and clinical course of tumors. The common steps of pathological analysis may be composed of biopsy, observation, microscopy, immunophenotyping, and electron microscopy observation. The radiological method is preoperative but the pathological one has to be postoperative.

For SRM patient management, options include aggressive treatment of radical nephrectomy and nephron-sparing methods such as active surveillance, partial nephrectomy, and ablative technologies [13].

1.3.2 Utility of CT images in Surgical Operation

CT, MRI, and ultrasound can all provide valuable information for characterizing renal masses and diagnosing them [14]. To characterize a solid renal mass, only MRI is able to replace CT scan in most circumstances while ultrasound is not adequate and out of date.

Due to the fact that MRI is still relatively new to clinical practice compared to CT and that a dataset of enough MRI images hasn't been established, there is little possibility that machine learning techniques could be utilized on MRIs. In contrast, four-phase CT has been clinical practice and standard modality for renal mass evaluation [15]. Thus, this study was carried out on CT images.

The scanning protocol of CT is composed of four stages, an unenhanced scan and contrast-enhanced acquisitions during the corticomedullary, nephrographic, and excretory stages. These four stage images are taken in order at different phases of illness. Each stage is a consecutive axial CT images of the tumor slice by slice. Slices of the first stage has a space of 5 mm while the other three has a space of 1 mm thus the number of latter three stages are always several times that of unenhanced images.

1.3.3 Related Researches on Algorithms Based Tumor Diagnosis

In the past decade, with the development of computer technology and radiography and their rapid infiltration into the medical field, the medical image processing field has witnessed explosive development. High-throughput quantitative feature extraction

converts those images into mineable data. These radiological data, together with other patient data, are analyzed by multiple models and algorithms, in order to improve diagnostic and predictive accuracy and to facilitate better clinical decision making, especially to treat patients with cancer [16].

Various algorithms have been proposed and implemented for fp-AML/RCC classification including analysis of CT histogram [17][18][19], enhancement patterns on contrast enhanced CT [20][21][22], etc. However, the results of these studies were inconsistent and conflicting.

Machine learning algorithms are among those algorithms that have been revolutionizing and promising in the medical image analysis field. Pattern classification has been utilized since 2008 when abnormalities were detected and characterized in chest radiographs using features of local image appearance [23]. The advancement of computer hardware made it possible to train more complex models on larger database. In recent several years, tumor classification using machine learning algorithms have got great momentum as they have been used in brain tumor classification, lung cancer differentiation, breast tumor categorization, urinary bladder cancer staging, etc [24][25][26][27][28].

As for renal masses, classification and regression tree (CART), as a branch of machine learning, has been used to differentiating AML from non-clear cell renal cell carcinoma (nccRCC) in [29] and logistic regression, k nearest neighbors (kNN), support vector machine (SVM) and random forest were utilized for classification of fp-AML and ccRCC in [30]. SVM was also used for categorization of fp-AML and RCC in [31][32]. There is a growing trend of applying deep learning algorithms in tumor image classification. CNN is the most typical and successful deep network in image processing. Originated from 1980, it was first used for medical image processing in 1995. It did not gain momentum until the emerging of efficient training networks and other advances. It is especially in last three years that deep learning was used for multiple kinds of medical image processing. Out of 47 papers published in 2015-2017 which focused on the problem of medical image classification, 36 are based on CNNs [33]. The work done in [34] is of great importance to the work in this paper as it is also classification using CNN

and four-phase CT images and proved CNN classifier to be qualified. The difference lies in that the former classifier applicable to liver masses while my work is intended for renal masses. There hasn't been much occurrence of papers featuring deep learning algorithm based renal masses classifiers till now and part of my work filled this vacancy.

1.4 Significance of the Research

As mentioned in 1.1, the classification of renal masses, with machine learning especially deep learning methods, into benign and malignant and further into three subdivisions of RCC is still an immature research field to be uncovered. The difficulty in collecting valid data set great obstacle to the research progress. On one hand, it lies in that there is always inadequate patient data in one hospital and the cooperation and sharing of data has to gone through certain procedures. On the other hand, the upgrading of medical image acquisition equipment hindered the collection of data of the same modality. In this research, data were collected from Changhai Hospital Affiliated to Second Military Medical University of Chinese PLA, one of the grade III-A general hospital in China (the highest level hospitals). A large amount of research data created favorable conditions for applying machine learning algorithms especially deep learning algorithm since its performance is just proportional to the amount of data fed to the system [35]. In contrast, the pathological examination of one patient would take a couple of hours and require specialized instruments. This is meaningful work and can be seen as a tentative, exploratory, and pioneering study and is expected to initiate a research wave of utilizing machine learning algorithms in kidney tumor image processing to alleviate doctors' burden and optimize hospital resource allocation, as well as reduce the cost of kidney tumor patients and reform treatment procedures.

1.5 Objective and Highlight of the Research

The goal of my work is to exploit SVM and CNN algorithms, which are typical

machine learning and deep learning methods that have been proved to be capable of medical image classification, to categorize SRMs into benign and malignant types and differentiate malignant tumors into pRCC, chRCC, and ccRCC. Effort has also been spared on CNN classifier with image augmentation However, it turned out that this approach did not show much promising and satisfying outcomes. Thus, it is not the focus of this study.

The research in this paper highlights the interaction between two disciplines which are radiology and computer science and a successful solution to the SRM classification problem. The work in this paper is completed in the hope that the utilization of machine learning especially deep learning algorithms in renal mass classification could cast some new light to the conventional diagnosis status quo and inspire more research fellows to take advantage of interdisciplinary cooperation, catch up with the rising tendency of machine learning and deep learning applications and come up with innovative, creative and impressive approaches to tackle practical problems. Since these are the four most common types of renal masses accounting for more than 90% of all renal tumors, this can be a relatively comprehensive classifier applicable to most of the patients.

1.6 Structure of the Research

This paper is composed of seven chapters. My thesis topic, background of the research, SRMs and current diagnosis techniques are briefed in this chapter and at the same time the goal, highlight, and significance of my work, which is to apply typical machine learning and deep learning algorithms, specifically SVM and CNN, for SRM CT images classification, are also described in Chapter 1.

Chapter 2 gives a thorough introduction of the source and modality of CT images and the preprocessing of them, including the segmentation of region of interest, background removal, and resize functions. The experimental design is also in this chapter which discussed the distribution of images into training, validation, and testing subsets.

Feature extraction along with the design and implementation of SVM classifier is

included in Chapter 3 while the establishment and application of CNN model is in Chapter 4. Results of both models showed in multiple meaningful parameters, together with comparison of their performance, are presented in Chapter 5.

Chapter 6 contains a concise conclusion followed by an in-depth discussion on a comprehensive evaluation of my work, the strengths and weaknesses, and future improvement directions.

Chapter 2

CT Images Collection, Preprocessing, and Experimental Design

2.1 Introduction

This chapter is to elaborate data source, collection and preprocessing as a preparation for the algorithm input. The preprocessing mainly includes four stages: selection, segmentation, background removal, and resize.

2.2 Data Used in This Paper

The data collection was approved by the Institutional Review Board of Shanghai Changhai Hospital Ethics Committee. The pathological analysis of all the patients were all performed by professional pathologists with rich experience and every diagnosis was confirmed by renal surgical resection specimens and Immunohistochemically (IHC) reports from January 2013 to July 2017. A total number of 1063 patients (169 AML, 710 ccRCC, 65 pRCC, and 119 chRCC) diagnosed to have developed AML or RCC were recorded. Following four limitations: 1) tumor type should be one of AML/ccRCC/pRCC/chRCC; 2) patient should have gone through four-phase CT renal

protocol acquisition; 3) the maximum SRM diameter should be less than 4.0 cm; 4) classic AML with detectable fat should be excluded since various conventional methods are capable of this, 863 cases were excluded and the left 200 patients of 200 renal masses (42 fp-AML, 97ccRCC, 26pRCC, and 35chRCC) were regarded to be eligible for this research. The detailed screening process is shown in Figure 2.1.

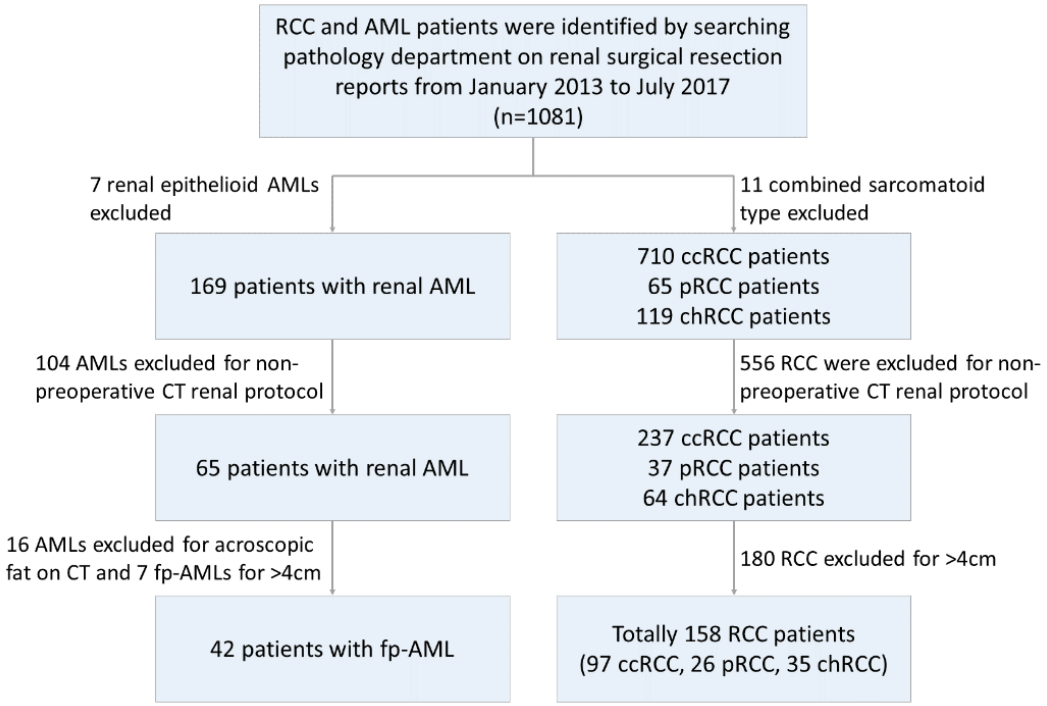

Figure 2.1: The screening process for establishing eligible dataset

The 200 cases consist of 76 females and 124 males whose age is within the range of 20-83 years and with a mean age of 55.07 years. Other patient characteristic related information can be found in Table 2.1.

Table 2.1: Patient characteristic information

Characteristic	RCC (158 cases)	fp-AML (42 cases)
Age (years)	56.18 ± 12.34	50.88 ± 9.37
Sex		
Male	113	11
Female	45	31
Lesion diameter (cm)	2.03 ± 0.78	2.28 ± 0.76
Side		
Right	86	18
Left	72	24

Age and lesion diameter data are mean ± standard deviation

In terms of image amount, there are 12317 images of the 200 SRMs in total. 786 first phase images were discarded for that the scale is too small to be fed to deep learning algorithm. Detailed numbers are in Table 2.2. It is found out that the first CT acquisition phase did not work well due to an obviously smaller number of images and artificially outlined contours with larger deviation compared to those of the other three phases. Therefore, the images of the first phase were all discarded in the final experiment.

Table 2.2: Image numbers of four categories and different phases

Stage	fp-AML (42 cases)	pRCC (26 cases)	chRCC (35 cases)	ccRCC (97 cases)
Corticomedullary	1027	584	744	1499
Nephrographic	1009	584	760	1547
Excretory	1008	570	699	1500
Contrast Enhanced	3044	1738	2203	4546

2.3 Segmentation, Background Removal, Resize, and Image Augmentation

All the patient images were acquired from Picture Archiving and Communication Systems (PACS) of Changhai Hospital and they are in the same format of Digital Imaging and Communications in Medicine (DICOM) and size of 512 px * 512 px, as is shown at the far left in Figure 2.2. DICOM images demand certain software for viewing. They were loaded and preprocessed by MATLAB 2015b. The lesion site was then manually segmented by labeling the contour of the renal mass and forming a closed pixel district. Whole-volume tumor segmentation, to ensure the reproductivity and quality [36][37], was entirely performed manually by a board-eligible radiologist with 5 years' experience who had no knowledge of the final lesion diagnosis during segmentation operation. In view of the fact that unenhanced images have slightly blurred boundaries, their segmentation was done by referring to the corresponding contrast-enhanced images. The borderlines were agreed upon by one other 10 year experienced radiologist.

The background was removed afterward as it cannot provide useful information to our solution. The last step of preprocessing was cropping the lesion site out of the large image onto a 96 px * 96 px image in which the tumor is in the center of the new image. This size was observed and proved to be a relatively small one that is capable of containing every lesion site. The resize function is of great help to shorten the algorithm runtime since the runtime is proportional to the image size (the size in the picture is not exactly the real size and it is just for demonstration).

Figure 2.2: Preprocessing of the original tumor image

For CNN classifier with image augmentation, new images were created by rotation, skew tilting, shearing, and random distortion functions. The acquired images were placed into appropriate folders waiting for training, validation and testing.

For each classifier, there are four folders three of which each includes images of three contrast enhanced phase, corticomedullary, nephrographic, and excretory phase. The last folder contains all the images of three phases. Correspondingly, four distinctive models were built, respectively called Model Corticomedullary, Model Nephrographic, Model Excretory and Model Enhanced.

2.4 Design of Experiment

SVM classifier works on extracted diagnosis features. Every patient has an average volume of 19 and a minimum of 8 axial scanned images for one of the three enhanced phases. The slices in the middle mostly comprise the largest cross sections while the uppermost and bottom ones are always too small for effective feature extraction. At the same time, the feature extraction operation of these images would cost a significant amount of time. Therefore, the seven images with the largest cross sections were picked out for each phase and a total of 21 images for each patient were used as SVM input.

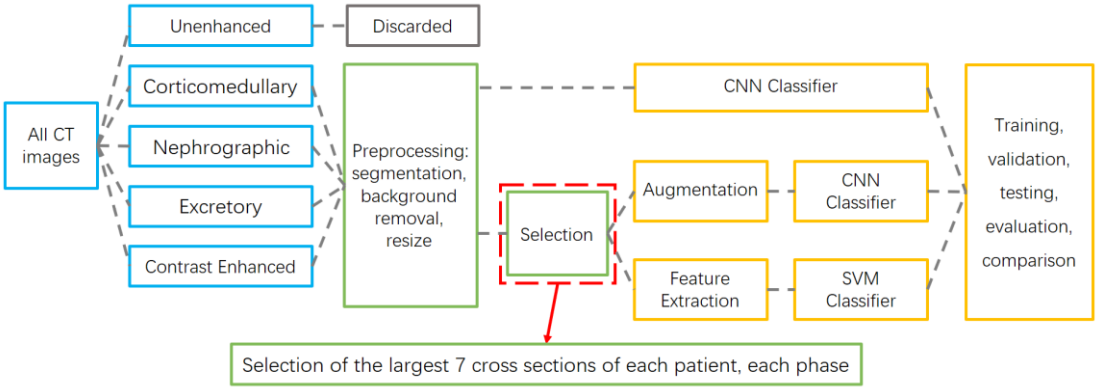

Figure 2.3: Dataflow of preprocessing and three classifiers

On the contrast, CNN always gain better performance when the dataset is larger. Thus

all available images were treated as input for CNN classifier. Furthermore, for the CNN classifier with image augmentation, image augmentation was executed on the same input as SVM. The dataflow is illustrated in Figure 2.3.

The next step was to place all the images into training, validation, and testing dataset. In practice, a certain combination of training and testing dataset will lead to a certain result that can be easily changed by altering the input dataset. To ensure that the performance evaluation of both SVM and CNN classifiers are objective and representative, and also to improve the generalization capacity of the classifiers, 20% of all cases were randomly picked out for test and the rest undertook 10-fold cross validation as is illustrated in Figure 2.4. During 10-fold cross validation, the overall non-testing subset is shuffled first and then divided into 10 distinct subsets randomly, each including 1/10 of the overall data. The ratios of benign/malignant and pRCC/chRCC/ccRCC were also almost the same (considering that number of images in some categories isn't divisible by 10).

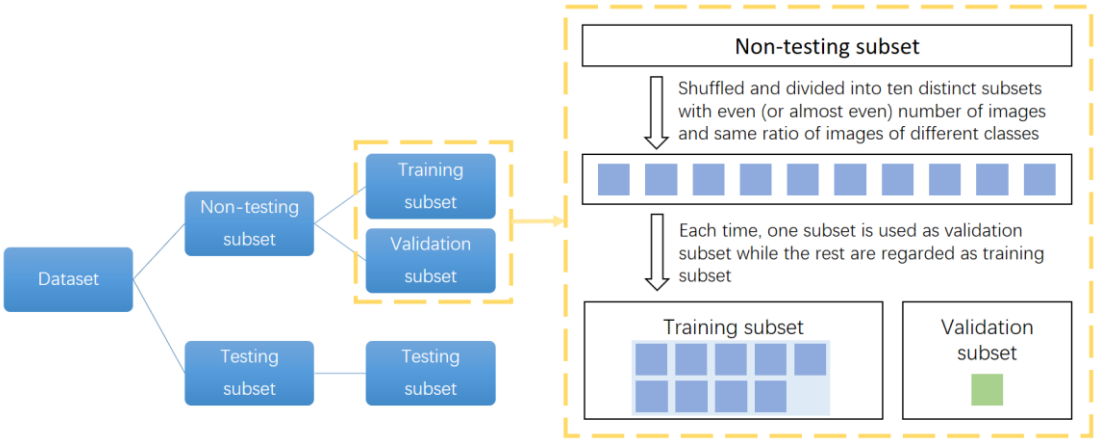

Figure 2.4: Distribution of training, validation, and testing subsets

The above-mentioned operations were implemented for both benign/malignant and pRCC/chRCC/ccRCC categorization. In the process of picking out test cases, the number calculated through total number multiplied by 20% was always not integer. In these situations, the numbers were rounded to the nearest greater number. For CNN classifier,

all the images should be loaded into it and the exact number of patients and images is presented in Table 2.3.

Table 2.3: Number of images fed to the CNN classifier

		fp-AML/RCC		pRCC/chRCC/ccRCC		
		fp-AML	RCC	pRCC	chRCC	ccRCC
Model Cortico-medullary	Total	42/1027	158/2827	26/584	35/744	97/1499
	Train	33/817	126/2239	20/458	28/576	77/1197
	Test	9/210	32/588	6/126	7/168	20/302
Model Nephro-graphic	Total	42/1009	158/2891	26/584	35/760	97/1547
	Train	33/801	126/2293	20/457	28/588	77/1241
	Test	9/208	32/598	6/127	7/172	20/306
Model Excretory	Total	42/1008	158/2769	26/570	35/699	97/1500
	Train	33/794	126/2211	20/454	28/556	77/1192
	Test	9/214	32/558	6/116	7/143	20/308
Model Enhanced	Total	42/3044	158/8487	26/1738	35/2203	97/4546
	Train	33/2412	126/6743	20/1369	28/1720	77/3630
	Test	9/632	32/1744	6/369	7/483	20/916

In 42/1027, 42 is the number of patients and 1027 is the number of images

The same patients were selected as testing cases and training cases for SVM classifier and CNN classifier with image augmentation. The detailed amounts of images fed to SVM classifier is demonstrated in Table 2.4. 7 images were selected for each phase and 21 for each patient. The numbers in Table 2.4 are applicable to Model Corticomedullary, Model Nephrographic, and Model Excretory. For Model Enhanced, the value should be three times of those in this table.

Table 2.4: Number of images fed to the SVM classifier

	fp-AML/RCC		pRCC/chRCC/ccRCC		
	fp-AML	RCC	pRCC	chRCC	ccRCC
Total	42 (294)	158 (1106)	26 (182)	35 (245)	97 (679)
Train	33 (231)	126 (882)	20 (140)	28 (196)	77 (539)
Test	9 (63)	32 (224)	6 (42)	7 (49)	20 (140)

As for CNN classifier, the testing subset is the same as CNN classifier while the training set of former three models were augmented to 10,000 and Model Enhanced contains 30,000 images in all.

2.5 Summary

In this chapter, the source, amount, modality, and characteristic of patient data are covered and the preprocessing including segmentation, background removal, and resize function is explained.

Chapter 3

SVM Method in SRM Classification

3.1 Introduction

As mentioned in 1.3.3, SVM, as an important embranchment of machine learning, has been seen in pervasive application in almost every field including medical image processing. In this chapter, basic principles of SVM, important diagnosis features extraction, establishment of SVM classifier are elaborated in sequence.

3.2 Basic Principles of SVM

Suppose there are a series of data of two classes as is illustrated in Figure 3.1 and the goal is to linearly separate them with a hyperplane, i.e. decision boundary, which is described in Equation (1):

$$\mathbf{w}^T\mathbf{x}+b=0 \tag{1}$$

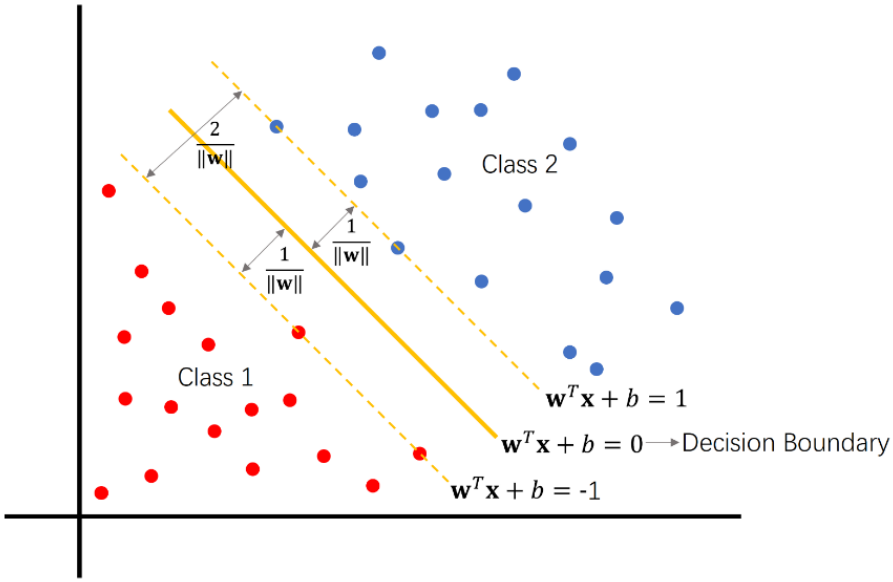

Figure 3.1: Decision boundary of SVM classification

Those blue dots above the decision boundary should have label 1 while those red dots below the hyperplane have label -1. Thus, this labelling scheme can be condensed to the formulation of $f(\mathbf{x})=sign(\mathbf{w}^T\mathbf{x}+b)$. If $f(\mathbf{x})>0$, then label should be 1 and if $f(\mathbf{x})<0$, then label should be -1.

In order to elevate the generalization ability, the best decision boundary should be the one that the distance between the two classes is the largest. SVM satisfy this requirement by maximizing the gap between the dots of different classes nearest to the hyperplane. Two other hyperplanes, $\mathbf{w}^T\mathbf{x}+b=1$ and $\mathbf{w}^T\mathbf{x}+b=-1$ are used to describe the dots nearest to the boundary. The goal has been transformed into maximizing the interval between these two hyperplanes, that is to say, the target is to maximize $2/\|\mathbf{w}\|$. It is equal to minimizing $\|\mathbf{w}\|$. To simplify the derivation and other calculations, it can be further modified into minimizing $\|\mathbf{w}\|^2/2$.

If the hyperplane still works perfectly for classification, then:

$$\begin{cases}\mathbf{w}^T\mathbf{x}_i+b\geq+1, & y_i=+1\\ \mathbf{w}^T\mathbf{x}_i+b\leq-1, & y_i=-1\end{cases} \tag{2}$$

(2) can be rewritten as:

$$y_i\left(\mathbf{w}^T\mathbf{x}_i+b\right)\geq 1 \tag{3}$$

(3) serves as the bound for objective function. So the purpose is to solve a convex quadratic programming problem as stated in (4):

$$\begin{aligned}&\min \quad \frac{1}{2}\|\mathbf{w}\|^2\\ &\text{subject to } \ y_i\left[\left(\mathbf{w}\mathbf{x}_i\right)+b\right]-1\geq 0\ (i=1,2,...,l)\left(l \textit{ is the sample number}\right)\end{aligned} \tag{4}$$

To solve the optimization problem in (4), Lagrangian function is constructed as:

$$\mathrm{L}\left(\mathbf{w},b,\alpha\right)\equiv\frac{1}{2}\mathbf{w}^T\mathbf{w}-\sum_{i=1}^{N}\alpha_i y_i\left(\mathbf{w}\mathbf{x}_i-b\right) \tag{5}$$

Take the derivatives of (5) with respect to $\mathbf{w}$ and b, then we have:

$$\begin{aligned}&\mathbf{w}=\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i\\ &\sum_{i=1}^{N}\alpha_i y_i=0\end{aligned} \tag{6}$$

Substitute (6) into (5), we get:

$$\begin{aligned}&\max. \quad W\left(\alpha\right)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1,\,j=1}^{N}\alpha_i\alpha_j y_i y_j\mathbf{x}_i^T\mathbf{x}_j\\ &\text{subject to } \ \alpha_i\geq 0,\sum_{i=1}^{N}\alpha_i y_i=0\end{aligned} \tag{7}$$

Thus, when we want to classify a dot, we only have to substitute it into:

$$\mathrm{g}\left(x\right)=\sum_{i=1}^{N}\alpha_i y_i\left\langle\mathbf{x}_i,\mathbf{x}\right\rangle+b \tag{8}$$

At this time, this hyperplane is capable of linear categorization which hardly tolerates any fault. To deal with nonlinear classification problems, kernel functions are of vital importance. Kernel functions serve the purpose by mapping low-dimensional samples to high-dimensional space so that they become linear separable. The classification function turns into (9):

$$\mathrm{h}\left(x\right)=\sum_{i=1}^{N}\alpha_i y_i\kappa\left(x_1,x_2\right)+b \tag{9}$$

where κ is the kernel function.

Due to the existence of outliers, a fault-tolerant SVM allows deviation of some dots from the decision boundary. Then the constraint can be represented as:

$$y_i\left(w^T x_i+b\right)\geq 1-\xi_i \tag{10}$$

where $\xi_i>0$ and ξ_i stands for slack variable. A cost function is introduced to make sure the fault-tolerance is within a certain limit, as is shown in (11):

$$\sum_{i=1}^{N} \xi_i \quad or \quad \sum_{i=1}^{N} \xi_i^2 \tag{11}$$

A cost coefficient C is used to represent the importance of the loss caused by outliers. The optimization function finally transforms into (12):

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & yi\left[wxi + b\right] \geq 1 - \xi_i \ \left(i = 1, 2, ..., l\right) \\ & \xi_i > 0 \end{aligned} \tag{12}$$

Construct the Lagrangian function as before then we get:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1,\, j=1}^{N} \alpha_i \alpha_j y_i y_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle \\ \text{subject to} \quad & 0 \leq \alpha \leq C, i = 1, 2, ..., n \\ & \sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned} \tag{13}$$

Eventually, this is a SVM capable of handling problems in practice.

SVM is intended for two-class differentiation problems. For multi-label classification, there exist several different techniques based on the traditional two-class categorization method. The most commonly used one in practice is one-versus-rest (OvR) classification. For an N-class classification problem, to employ OvR method is to build N one-versus-rest classifiers and to choose the class which classifies the test data with the largest margin. Another strategy is one-versus-one (OvO) method which involves the establishment of $\mathrm{N(N-1)/2}$ classifiers and selection of the class that is chosen by the most classifiers.

3.3 Extraction of Important Diagnostic Features

In this study, four types of diagnostic features were extracted, histogram, shape, texture, and wavelet features. These four categories of features are among the most representative features of images. Histogram features comprise non-zero maximum and minimum values of pixel gray level, mean and median of non-zero pixel gray levels, and six other histogram features including entropy, mean, variance, skewness, kurtosis, and uniformity. Shape features constitute the major and minor axis, orientation, overall area, perimeter, solidity, and eccentricity of each SRM. Texture features mainly include gray

level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size zone matrix (GLSZM), neighborhood gray tone difference matrix (NGTDM). The last kind is two-dimensional wavelet features. Feature extraction process was also completed in MATLAB 2015b.

3.4 SVM Classification

After making several attempts on different kernels, it turned out that linear kernel has the best performance in this study while another popular one, RBF (Radial Basis Function) kernel, performed poorly on this dataset. For pRCC/chRCC/ccRCC differentiation, OvR approach was executed. The training, validation, and testing dataset allocation has been articulated in 2.4. The classification was carried out by using Python 3.5.2.

3.5 Summary

In this chapter, the basic principles of SVM and the design of the SVM classifier applicable for this study are introduced. Meanwhile, the feature extraction for each image is also included in this part.

Chapter 4

CNN Method in SRM Classification

4.1 Introduction

CNN, as a deep neural network, has been amazingly useful in multiple fields including medical image categorization. However, no endeavor has been made in SRM classification and that's where my work comes in. In this chapter, the basic principles of CNN, the development of a specific CNN model, and the image augmentation operations are covered. Additionally, the training, validation, and testing of the CNN classifier is also stated in this chapter.

4.2 Basic Principles of CNN

A CNN is composed of at least three layers with the first layer as the input layer, the last layer as the output layer and all the rest as hidden layers. The structure of a CNN with two hidden layers is demonstrated in Figure 4.1.

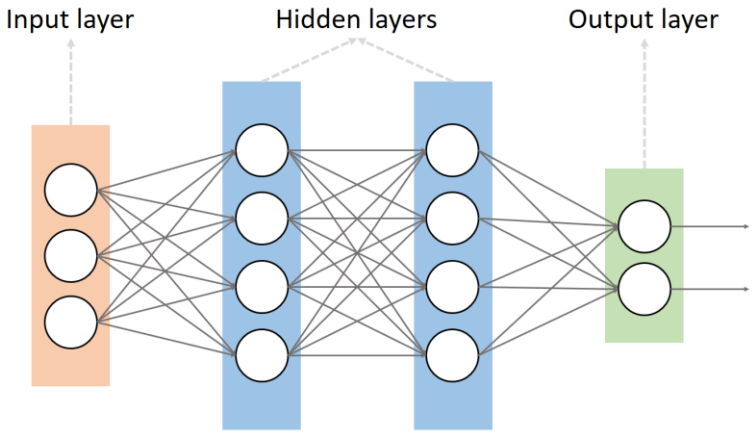

Figure 4.1: Structure of a CNN with two hidden layers

Each circle in Figure 4.1 is a neuron, which is the basic computing unit in a CNN. For activation layer, when several inputs are fed to a neuron, every input will be multiplied by a weight and after that the sum will be added a bias value. The total is used as the input of the activation function whose output is also the output of the neuron. The operation of a neuron is depicted in Figure 4.2. The weights can be initialized by all-zeros or random values and will be updated during the training process. Activation functions used in this study are Sigmoid and rectified linear unit (ReLU). Its function lies in that it manages to add nonlinear elements to a linear model which is greatly limited to tackle practical issues. The activation function of Sigmoid is described as follows:

$$y = 1/\left(1+e^{-z}\right) \tag{14}$$

For ReLU activation function, it is expressed as:

$$y = \begin{cases} 0, & \text{if } (x<0) \\ x, & \text{if } (x\geq 0) \end{cases} \tag{15}$$

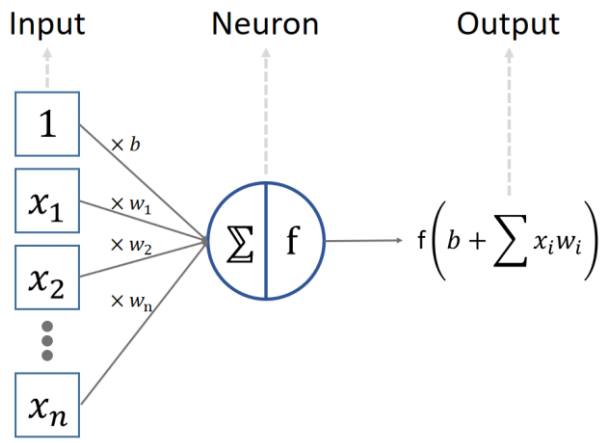

Figure 4.2: Operation of a neuron

Convolutional layers use a preset convolution kernel to perform a convolution operation on the original image pixel matrix. It is applied to filter signal and find patterns in the signal. The front convolutional layers extract low-level features and the latter convolutional layers extract higher level features. The convolutional calculation method is shown in Figure 4.3.

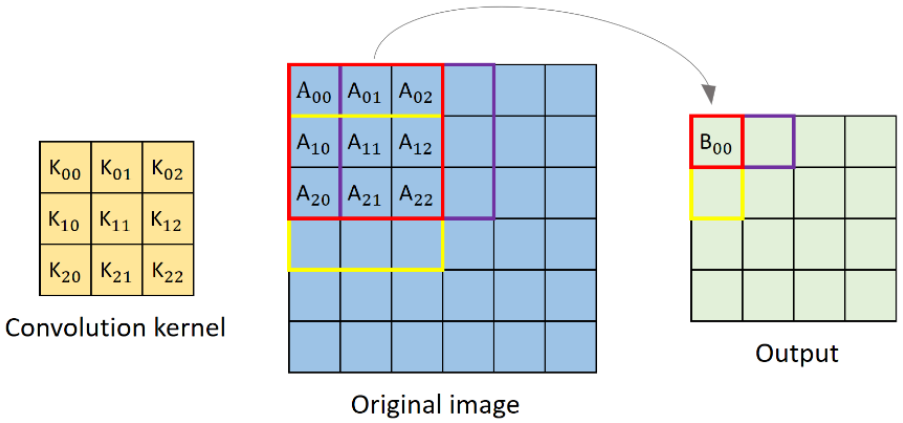

Figure 4.3: Convolutional calculation method

For a 6*6 pixel original image, if a 3*3 convolution kernel is used, the kernel will perform a convolution operation on the part of the original image in the red box and the result will be placed in the small red box in the output matrix. Then the convolution kernel will move to the right purple box and perform a convolution operation and the result will

be placed in the box with the corresponding color. The convolution kernel will continue to move to the right until reaching the end of this row and then follows the same operations on the next row starting from the leftmost yellow box till arriving at the bottom right corner. The convolution operation follows equation (16):

$$B(i,j)=\sum_{m=0}\sum_{n=0}K(m,n)*A(i-m,j-n) \tag{16}$$

where A is one part of the original image like the red box, K is the convolution kernel and B is the output. Specifically, in Figure 4.3, there exists:

$$\begin{aligned} B_{00}=&A_{00}K_{00}+A_{01}K_{01}+A_{02}K_{02}\\ &+A_{10}K_{10}+A_{11}K_{11}+A_{12}K_{12}\\ &+A_{20}K_{20}+A_{21}K_{21}+A_{22}K_{22} \end{aligned} \tag{17}$$

Pooling layer is always immediately after the convolutional layer and mainly falls into two categories, max-pooling layer and mean-pooling layer. Taking max-pooling layer as an example, if the stride is 2, then the maximum value of every 2*2 pixel square in the original image is used as the output corresponding to the pooled position as is illustrated in Figure 4.4. After a pooling layer, the number of rows and columns will be respectively cut down to 1/2 of the previous value, which is of significant importance in decreasing the spatial resolution of the input matrix and sharply reduce operation time.

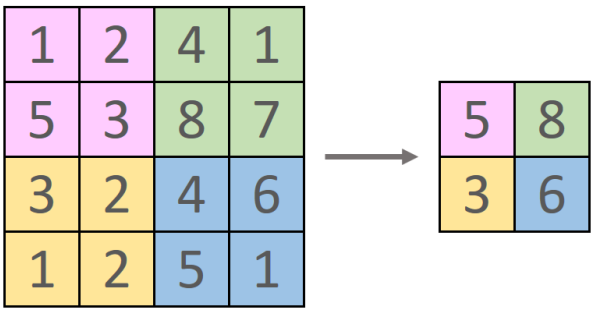

Figure 4.4: Operation of a max-pooling layer

Flatten layer's role lies in converting a multi-dimensional matrix into a one-dimensional vector. It is always used before fully connected layers. Fully connected layer got its name because each neuron in a fully connected layer has an input of all the neurons of the upper layer. Possibilities of one case to fall into each class will be calculated with

a sum of 1 and the category with the largest possibility will be the decision class.

Since the amount of effective images in some fields such as medical images is often of very small scale, an overly complex network structure can easily lead to overfitting, that is to say, the network structure is excellent for this batch of input, but once testing with a new input, the performance will degrade dramatically and the model lacks the ability of generalization. To avoid this situation, dropout layers are added to CNNs. During the training process, for the input of an upper layer, a certain proportion of network parameters are randomly discarded in each epoch. In other words, a proportion of neurons are ignored in each epoch as is shown in Figure 4.5. In this figure, the dropout rate is 0.4 for all the three layers since 40% of all the neurons in each layer are ignored.

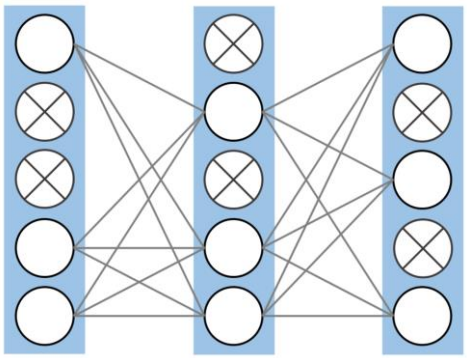

Figure 4.5: Illustration of dropout layer

This helps the classifier to learn more robust features and prevent it from overfitting. At the same time, as the number of parameters decreases, the speed of calculation will increase.

4.3 Establishment of CNN Classifier

The CNN classifier in this study is mainly composed of 12 layers, 4 convolutional layers each followed by a max-pooling layer, 1 flatten layer and 3 densely connected layers, as is shown in Figure 4.6. Dropout functions and activation functions also played an important role in this model.

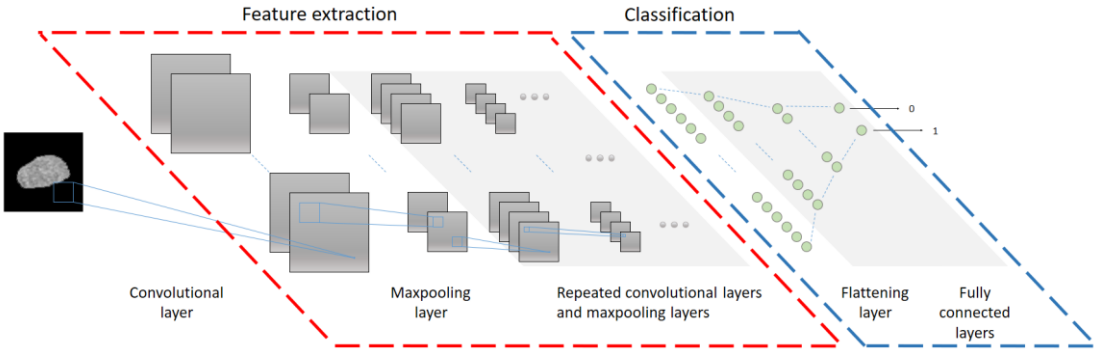

Figure 4.6: CNN structure used in this research

The convolution kernel used in this study is 5*5 and the stride of max-pooling layer is 2. Each pooling layer is followed by a ReLU activation function to add nonlinear elements. After converting data into one-dimensional vector, three full-connected layers downsize the number of neurons and finally give a decision class as an output. The fully connected layers except for the last one is respectively followed by a dropout function with a dropout ratio of 0.5. Sigmoid activation function comes after the last fully connected layer.

4.4 CNN Classifier with Image Augmentation

As deep learning is a data hungry network and considering the fact that an imbalanced dataset always contributes to inclination towards the larger group, image augmentation was taken to enlarge the dataset. New images were acquired by rotation to the left or right within a range of 10°, skew tilting with a largest magnitude parameter of 0.1, shearing to the left or right within a magnitude of 10, and random distortion within a magnitude of 2. These functions were executed under preset probabilities on the same dataset as in SVM. The effect of these operations is shown in Figure 4.7. The original image is a randomly selected tumor image. Since the tumor images contain complicated texture features, the augmentation methods represented by the newly gained images are not intuitive enough. Therefore, the smaller images with white and black squares are

added to demonstrate clearly the four types of image augmentation methods.

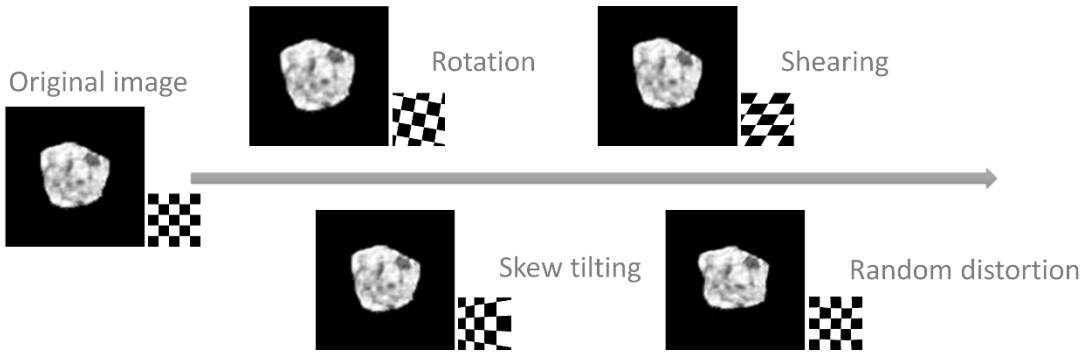

Figure 4.7: Image augmentation methods

By this way, all the three phases of different categorizations were augmented to a number of 10,000 images and Model Enhanced for fp-AML/RCC differentiation enjoys a dataset of 60,000 images. As this augmentation method did not show any improvement compared to conventional CNN, this method is deserted at last and it is not the emphasis of this study.

4.5 Training, Validation, and Testing of CNN

As is stated in 2.4, 10-fold cross validation with shuffling and stratified sampling was carried out on randomly selected 80% of the entire dataset and testing was conducted on the rest of the dataset the cases of which were randomly selected. The parameters were updated during each batch and epoch with input labels as teaching data. The parameters would remain unchanged when reaching the end of training process and would make predictions on the testing subset.

Four different models each led to 10 mutually independent training processes and 10 distinct classifiers as 10-fold cross validation was used. For fp-AML classification, each training process lasted 80 epochs with a batch size of 32. For pRCC/chRCC/ccRCC differentiation, the epoch number was preset as 120 since the parameters were upgraded much slowly. Each image in the training subset was fed to the classifier once every epoch.

At the beginning of each epoch, the images undertook batch normalization. In the training process, cross entropy was designed to be the lost function and the difference between predicted labels and true labels was transmitted to previous layers as a feedback, and the parameters of each layer were slowly updated. With the increase of training time, the loss trended down with a decreasing declining rate until reaching a point when it started fluctuating within a small limit. SGD optimization algorithm was applied. Since an excessive learning rate easily results in local optimal solutions and an inadequate learning rate lengthens the training process to a great extent, 0.01 was finally chosen as learning rate after adjusting a couple of times for fp-AML classification. For pRCC/chRCC/ccRCC categorization, the learning rate was 0.001 to escape falling into local minimal. Each training process ended with computing accuracy and confusion matrix.

The validation and testing process included label predictions and calculation of various parameters which will be elaborated in the next chapter.

The establishment of CNN model and CNN with image augmentation was also coded in Python 3.5.2 and it was based on an open source machine learning library, TensorFlow 1.3.0 and high level application programming interfaces (APIs) provided by Keras 2.0.9.

4.6 Summary

In this chapter, the basic knowledge of CNN, the establishment of a CNN model for this research and the image augmentation approach are covered. In addition, it gives a clear description of the training, validation, and testing process of the CNN classifier.

Chapter 5

Results

5.1 Introduction

This chapter presents the results of above introduced three algorithms, SVM, CNN, and CNN with image augmentaiton for fp-AML/RCC categorization and pRCC/chRCC/ccRCC categorization Each algorithm was extended to four Models, Model Corticomedullary, Model Nephrographic, Model Excretory, and Model Enhanced. Each model contains 10 classifiers as 10-fold cross validation was applied. Multiple parameters were calculated for a comprehensive and objective evaluation of the models. Comparison between these algorithms and models will also be covered in this chapter.

5.2 Model Performance Evaluation Methods

All the programs of modelling ran on a same computer, Dell XPS 8900, with an i7-6700 3.40 GHz Intel Core CPU and a 24 GB random access memory (RAM). The evaluation was based on APIs offered by scikit-learn 0.19.1.

In fp-AML/RCC classification, fp-AML was preset as class ‘0’ and RCC was preset as class ‘1’. True positive/negative refers to correctly classified RCCs/fp-AMLs while false negative represents RCC misclassified into the category of fp-AML and false

positive represents just the opposite. For each classifier in each fp-AML/RCC differentiation model, receiver operating characteristic (ROC) curve was drawn and multiple indicators were calculated, including area under curve (AUC), confusion matrix, accuracy, sensitivity, specificity, false positive rate, false negative rate, precision, and runtime. The mean ROC was also drawn for each model, as well as other above mentioned indicators. To examine the performance of each Model in real life circumstances, a decision classification of the patient was predicted depending on the prediction of all the images of each test patient in all the 10 classifiers of the model. For example, if a test patient case includes 10 images, 10 independent classifiers will yield 10 predictions on each image, thus there are 100 image classification decisions. Then the decision of the patient is made relying on the 100 decisions. A simple method is to follow the principle that the minority obeys the majority.

In pRCC/chRCC/ccRCC categorization, pRCC, chRCC, and ccRCC are respectively preset as class ‘0’, ‘1’, and ‘2’. While most of the above mentioned parameters can be only obtained for two-category classification problems, accuracy, confusion matrix of classifiers and decision of patients remained as assessment indicators.

CNN with image classification was applied in fp-AML/RCC classification. Since it did not show any improvement while demanding significantly longer time for training, it was discarded and wasn’t applied to pRCC/chRCC/ccRCC classification.

5.3 fp-AML/RCC Classification Results

This section presents all the results related to four models for fp-AML/RCC classification mainly in terms of ROC curve, multiple other evaluation indicators listed above, and patient diagnostic prediction. Comparison between these models using two different algorithms, SVM and CNN, is discussed at the end of this section.

5.3.1 Model Corticomedullary

The ROC curves of SVM and CNN in Model Corticomedullary are shown in Figure

5.1. a) is validation ROCs for SVM; b) is testing ROCs for SVM; c) is validation ROCs for CNN; d) is the testing ROCs for CNN. In each ROC image, the light-colored curves represent ROC of each classifier while the bold blue one demonstrate the mean ROC. The red dashed line is the reference line with a AUC of 0.5. In Model Corticomedullary, it is apparent that AUC of SVM is quite similar to that of CNN. However, SVM classifiers work much better than CNN in terms AUC. The AUC on testing subset is always worse than validation subset.

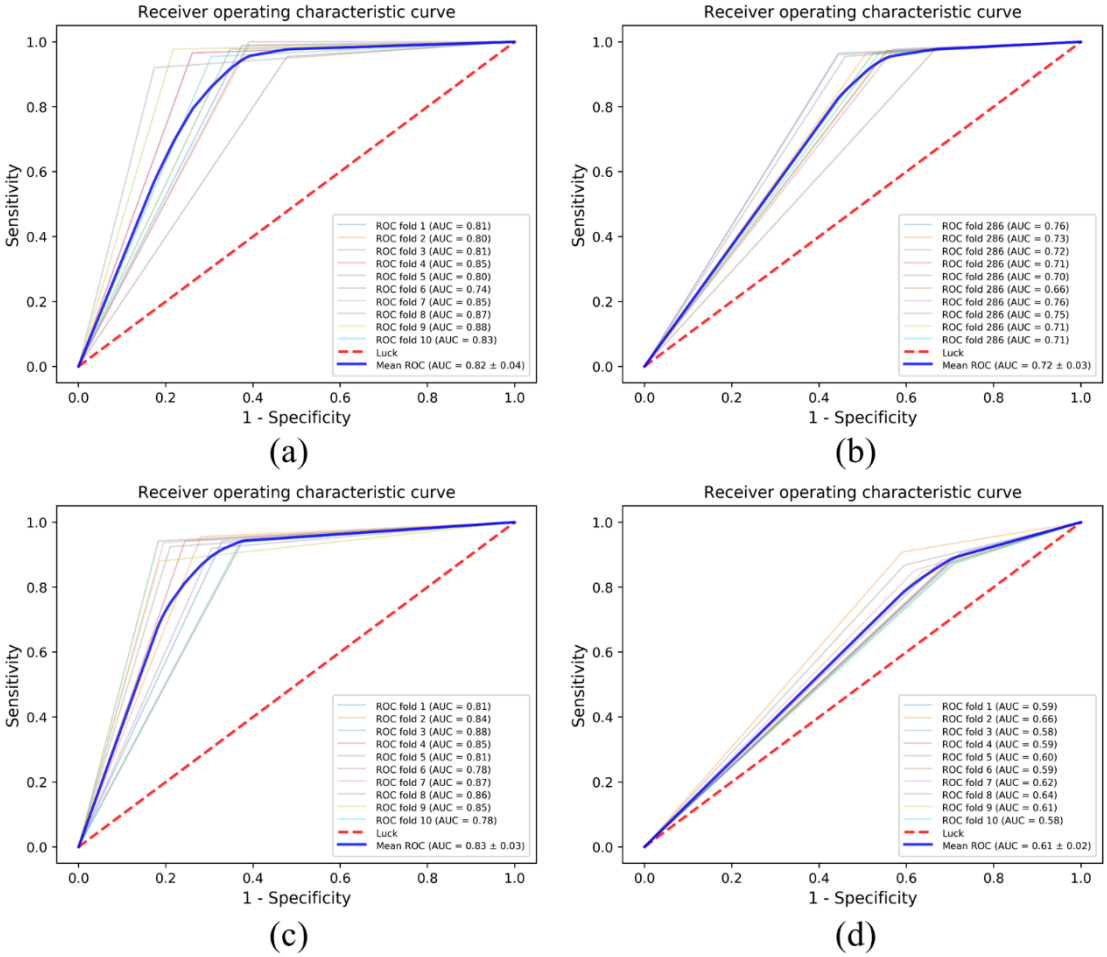

Figure 5.1: Model Corticomedullary: ROCs of SVM and CNN validation and testing

All the other parameters of validation and testing are presented in Table 5.1. As the performance assessment of each classifier in the same Model is not objective and

representative enough, only the mean values of the 10 classifiers are listed here. Patient class prediction 1 is the decision made following the above mentioned rule of minority obeying majority. If there are same numbers of predictions of two classes, then the decision class will be fp-AML since the model tends to misclassify fp-AML as RCC. Patient class prediction 2 is the decision made when the number of RCC predictions exceeds 70% of all images' predictions. Considering the fact that false positive rate is much higher than false negative rate, and the model leans to the side of RCC decision on the images, patient class prediction 2 works better than patient class prediction 1 by increasing specificity to a large extent while remaining a satisfying sensitivity.

Table 5.1: Performance evaluation of Model Corticomedullary

	SVM		CNN	
	Validation	Testing	Validation	Testing
Mean AUC	0.82	0.72	0.83	0.61
Accuracy	0.91	0.86	0.88	0.73
Sensitivity	0.96	0.97	0.93	0.88
Specificity	0.71	0.49	0.73	0.33
False positive rate	0.29	0.51	0.27	0.67
False negative rate	0.04	0.03	0.07	0.12
Precision	0.93	0.87	0.91	0.79
Patient class prediction 1				
Accuracy	-	36/41 (0.88)	-	32/41 (0.78)
Specificity	-	5/9 (0.56)	-	1/9 (0.11)
Sensitivity	-	31/32 (0.97)	-	31/32 (0.97)
Patient class prediction 2				
Accuracy	-	38/41 (0.93)	-	36/41 (0.88)
Specificity	-	7/9 (0.78)	-	6/9 (0.67)
Sensitivity	-	31/32 (0.97)	-	30/32 (0.94)

From Table 5.1 we can clearly see that SVM performs better than CNN algorithm

with Corticomedullary stage images in terms of multiple parameters including mean AUC, accuracy, sensitivity, specificity, false positive rate, precision, and patient class prediction 1 and 2. SVM also enjoys a smaller difference between validation and testing subset which means it has higher ability of generalization. SVM gains an accuracy of 0.91 and sensitivity of 0.96, which are quite satisfying compared to what previous works has achieved. In patient class prediction, it has correctly diagnosed 38 cases out of a total of 41, which indicates that it has the ability of pre-diagnosis in real life. The reason why patient class prediction 2 works better than patient class prediction 1 attributes to an imbalanced dataset. There are more RCC samples than fp-AMLs and it results in leaning to RCC decisions. CNN is more easily to fall into this problem. Most of the time, for patient class prediction, a patient with RCC has an accuracy of more than 70% while a patient with fp-AML has only an accuracy of less than 50%. We artificially balance the model by increasing the criteria of RCC decisions.

5.3.2 Model Nephrographic

The ROC curves of SVM and CNN in Model Nephrographic are presented in Figure 5.2. The images and colored curves have the same meaning as in Figure 5.1. It takes on same trends as the ROC curves in Model Corticomedullary that SVM has better AUCs.

The other parameters of validation and testing are presented in Table 5.2.

It is evident that SVM performs better than CNN by using nephrographic stage images in terms of most evaluation parameters. SVM has reached an accuracy of 0.85 on testing subset. Patient class prediction 1 reaches an accuracy of 0.85 and a very high sensitivity without sacrificing specificity too much. In medical practice, sensitivity is more significant than specificity since both doctors and patients don’t want to miss any malignant renal masses. Misdiagnosis of benign renal masses will only lead to excessive examination and treatment, while misdiagnosis of malignant ones may cause serious consequences and are not acceptable. Patient class prediction 2 increases specificity and decreases sensitivity at the same time, which is in line with the explanations as in Model Corticomedullary. CNN is kind of overfitting with images of nephrographic stage.

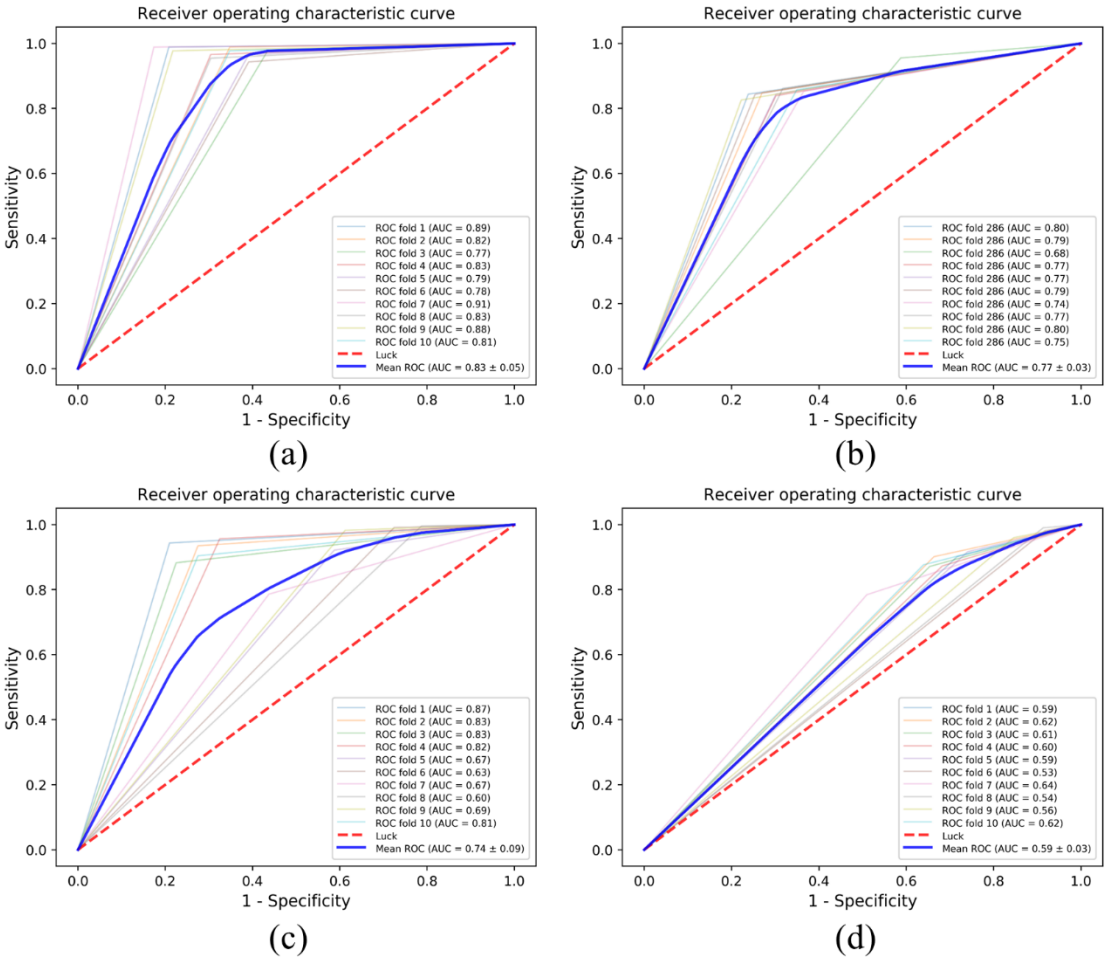

Figure 5.2: Model Nephrographic: ROCs of SVM and CNN validation and testing

Table 5.2: Performance evaluation of Model Nephrographic

	SVM		CNN	
	Validation	Testing	Validation	Testing
Mean AUC	0.83	0.77	0.74	0.59
Accuracy	0.91	0.81	0.83	0.74
Sensitivity	0.97	0.85	0.93	0.91
Specificity	0.71	0.69	0.55	0.27
False positive rate	0.29	0.31	0.45	0.73
False negative rate	0.03	0.15	0.07	0.09
Precision	0.93	0.81	0.86	0.78
Patient class prediction 1				
Accuracy	-	35/41 (0.85)	-	32/41 (0.78)
Specificity	-	6/9 (0.67)	-	0/9 (0.00)
Sensitivity	-	29/32 (0.91)	-	32/32 (1.00)
Patient class prediction 2				
Accuracy	-	33/41 (0.80)	-	32/41 (0.78)
Specificity	-	8/9 (0.89)	-	2/9 (0.22)
Sensitivity	-	25/32 (0.78)	-	30/32 (0.94)

5.3.3 Model Excretory

The ROC curves of SVM and CNN in Model Nephrographic are presented in Figure 5.3. The images and colored curves have the same meaning as in Figure 5.1. It takes on same trends as the ROC curves in Model Corticomedullary and Model Nephrographic that SVM has better AUCs.

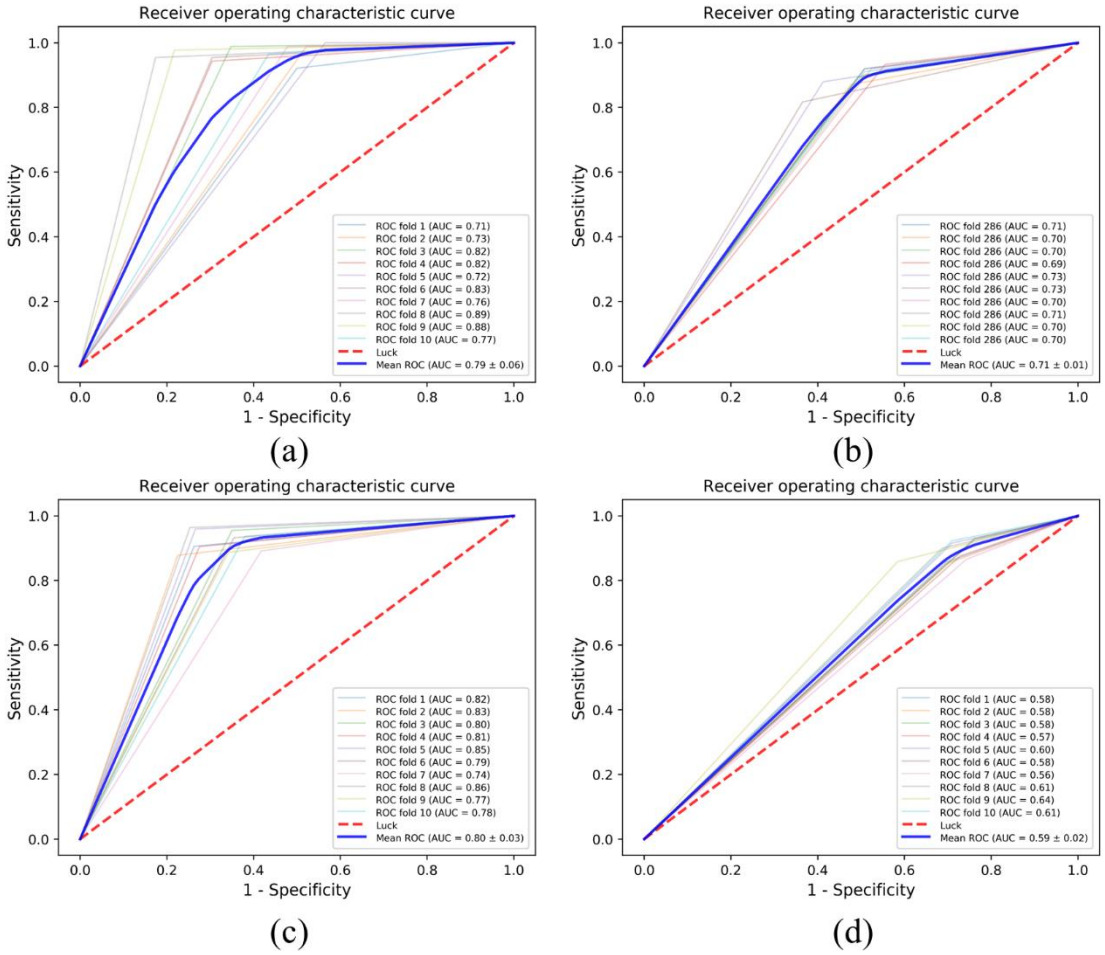

Figure 5.3: Model Excretory: ROCs of SVM and CNN validation and testing

The other parameters of validation and testing are presented in Table 5.3.

Table 5.3: Performance evaluation of Model Excretory

	SVM		CNN	
	Validation	Testing	Validation	Testing
Mean AUC	0.79	0.71	0.80	0.59
Accuracy	0.90	0.81	0.86	0.72
Sensitivity	0.97	0.89	0.92	0.89
Specificity	0.65	0.54	0.69	0.29
False positive rate	0.35	0.46	0.31	0.71
False negative rate	0.03	0.11	0.08	0.11
Precision	0.85	0.87	0.89	0.77
Patient class prediction 1				
Accuracy	-	34/41 (0.83)	-	32/41 (0.78)
Specificity	-	4/9 (0.44)	-	0/9 (0.00)
Sensitivity	-	30/32 (0.94)	-	32/32 (1.00)
Patient class prediction 2				
Accuracy	-	36/41 (0.88)	-	35/41 (0.85)
Specificity	-	6/9 (0.67)	-	4/9 (0.44)
Sensitivity	-	30/32 (0.94)	-	31/32 (0.97)

5.3.4 Model Enhanced

The ROC curves of SVM and CNN in Model Enhanced are presented in Figure 5.4. The images and colored curves have the same meaning as in Figure 5.1. It takes on same trends as the ROC curves in former three models with images of one contrast enhanced phase that SVM has better AUCs.

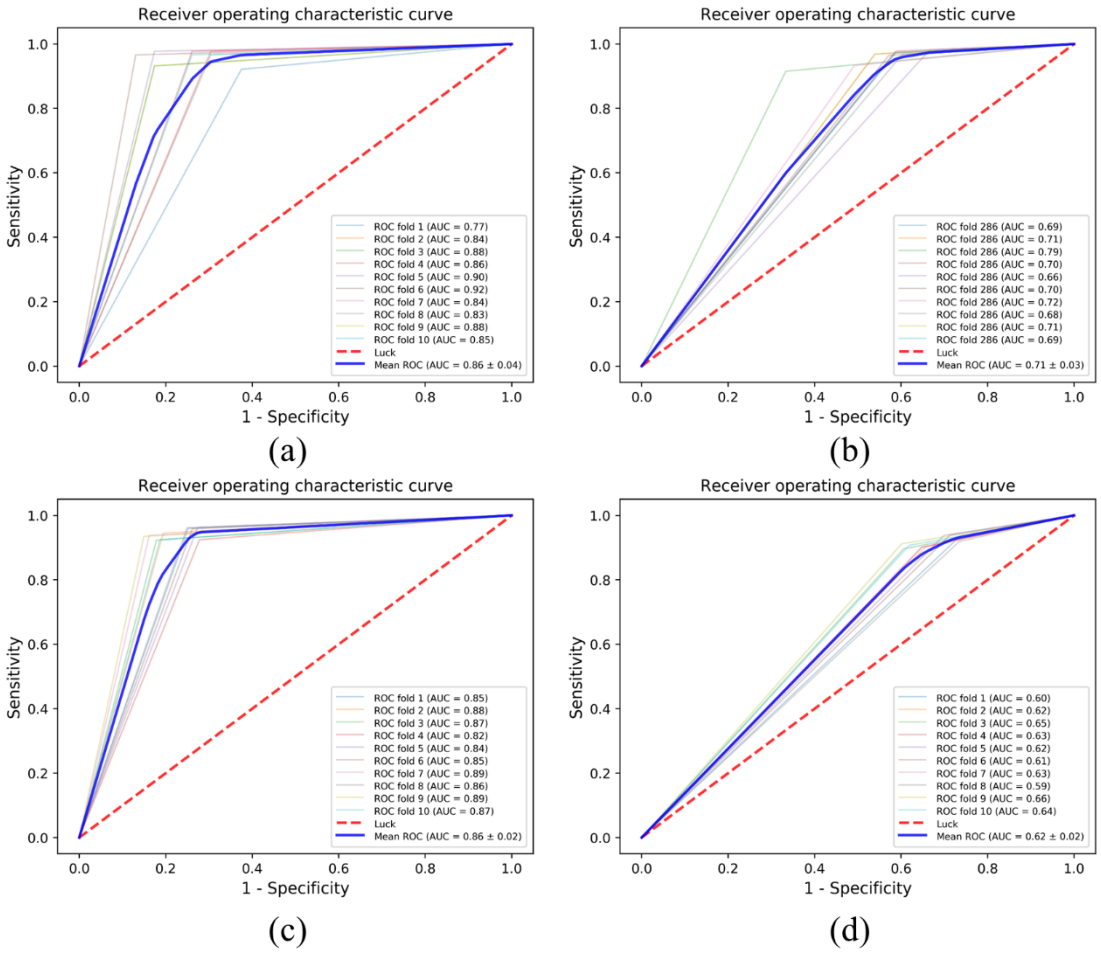

Figure 5.4: Model Enhanced: ROCs of SVM and CNN validation and testing

The other parameters of validation and testing are presented in Table 5.4.

Table 5.4: Performance evaluation of Model Enhanced

	SVM		CNN	
	Validation	Testing	Validation	Testing
Mean AUC	0.84	0.73	0.86	0.62
Accuracy	0.92	0.87	0.90	0.76
Sensitivity	0.97	0.99	0.94	0.91
Specificity	0.71	0.44	0.78	0.33
False positive rate	0.29	0.56	0.22	0.67
False negative rate	0.03	0.01	0.06	0.09
Precision	0.92	0.86	0.92	0.79
Patient class prediction 1				
Accuracy	-	36/41 (0.88)	-	32/41 (0.78)
Specificity	-	4/9 (0.44)	-	1/9 (0.11)
Sensitivity	-	32/32 (1.00)	-	32/32 (1.00)
Patient class prediction 2				
Accuracy	-	37/41 (0.90)	-	36/41 (0.88)
Specificity	-	6/9 (0.67)	-	4/9 (0.44)
Sensitivity	-	31/32 (0.97)	-	32/32 (1.00)

With all the contrast enhanced images as input, SVM is relatively more advantageous than CNN in testing. CNN has reached an accuracy of 0.88, slightly lower than that of SVM which has achieved 0.90. CNN is excellent in sensitivity but SVM is obviously better than CNN in specificity.

5.3.5 Comparison Among Different Models and Algorithms

By utilizing SVM algorithm, the mean ROC curve in testing for each model is illustrated in Figure 5.5. There is no much difference between Model Corticomedullary, Model Excretory and Model Enhanced. Model Nephrographic has the largest AUC. The comparison among other parameters for testing is shown in Table 5.5.

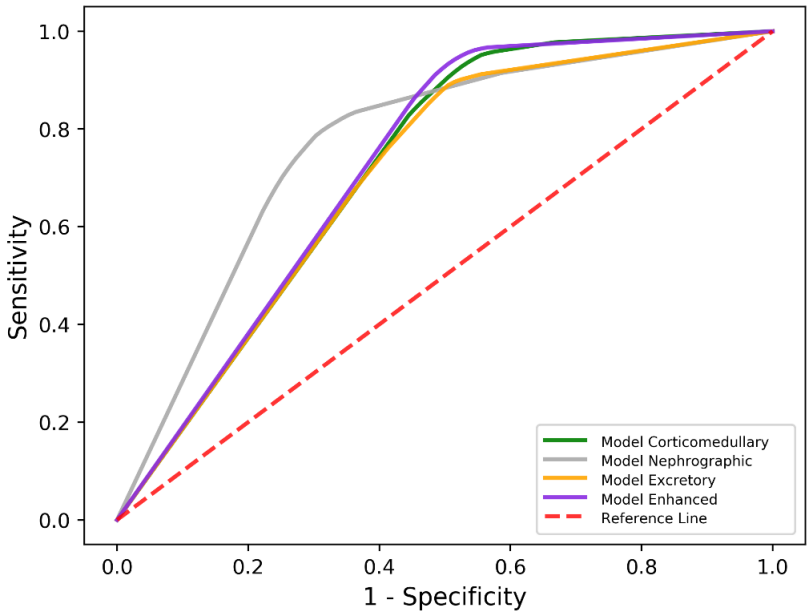

Figure 5.5: Comparison between ROC curves of four models using SVM

Model Corticomedullary possesses the best accuracy and patient class prediction 1 and 2. Model Nephrographic has the best AUC but it performs poorly in other evaluation items. For real-world patient diagnosis, Model Corticomedullary has achieved a total accuracy of 0.93 and a tremendous sensitivity of 0.97 and a basically satisfactory specificity of 0.78. As sensitivity is the most concerned assessment parameter in practice, this Model can be regarded as an initial success in application of algorithms in SRM classification instead of pathological examination. The Model Enhanced does not show any improvement on the basis of other three models although it has a larger scale of input.

Table 5.5: Comparison between performance evaluation of four models using SVM

	Model Cortico-medullary	Model Nephro-graphic	Model Excretory	Model Enhanced
Mean AUC	0.72	0.77	0.71	0.73
Accuracy	0.86	0.81	0.81	0.87
Sensitivity	0.97	0.85	0.89	0.99
Specificity	0.49	0.69	0.54	0.44
False positive rate	0.51	0.31	0.46	0.56
False negative rate	0.03	0.15	0.11	0.01
Precision	0.87	0.81	0.87	0.86
Patient class prediction 1				
Accuracy	36/41 (0.88)	35/41 (0.85)	34/41 (0.83)	36/41 (0.88)
Specificity	5/9 (0.56)	6/9 (0.67)	4/9 (0.44)	4/9 (0.44)
Sensitivity	31/32 (0.97)	29/32 (0.91)	30/32 (0.94)	32/32 (1.00)
Patient class prediction 2				
Accuracy	38/41 (0.93)	33/41 (0.80)	36/41 (0.88)	37/41 (0.90)
Specificity	7/9 (0.78)	8/9 (0.89)	6/9 (0.67)	6/9 (0.67)
Sensitivity	31/32 (0.97)	25/32 (0.78)	30/32 (0.94)	31/32 (0.97)

By applying SVM algorithm, the mean ROC curve in testing for each model is illustrated in Figure 5.6. The curves take on the same tendency and their AUC are within a small range. Model Enhanced has the largest AUC while Model Nephrographic possesses the smallest AUC.

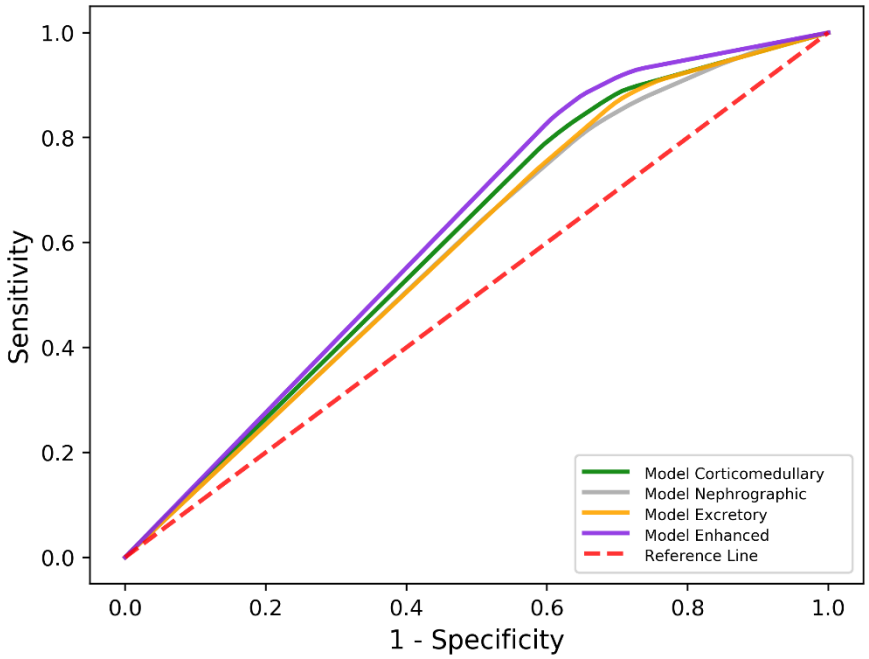

Figure 5.6: Comparison between ROC curves of four models using CNN

The comparison among other parameters for testing is shown in Table 5.6.

Model Enhanced obtained the best AUC, accuracy, sensitivity, specificity, and precision among the four models using CNN, which may result from a larger dataset. In real-world circumstances for patient diagnosis, Model Enhanced and Model Corticomedullary reached the same accuracy of 0.88. The four models all lean to RCC class as the number of RCC images is three times that of fp-AML images.

Feature extraction is the most time consuming process in SVM. It took 3.65 h to extract features of Corticomedullary stage images. The training and testing process in one fold respectively took 18.89 s and 0.07 s. It took less than 0.01 s to test on each patient, which is totally within the acceptable range. It took around 8.03 h for the 10-fold training and cross validation of Model Corticomedullary by using CNN. The testing process took 8 s for all patients and less than 0.2 s for each patient diagnosis. In real-world situations, reducing testing time is of higher priority and four models using SVM and CNN are all satisfactory in this concern as 0.01 s and 0.2 s are both fast enough and they don't have much difference in practical use.

Table 5.6: Comparison between performance evaluation of four models using CNN

	Model Cortico-medullary	Model Nephro-graphic	Model Excretory	Model Enhanced
Mean AUC	0.61	0.59	0.59	0.62
Accuracy	0.73	0.74	0.72	0.76
Sensitivity	0.88	0.91	0.89	0.91
Specificity	0.33	0.27	0.29	0.33
False positive rate	0.67	0.73	0.71	0.67
False negative rate	0.12	0.09	0.11	0.09
Precision	0.79	0.78	0.77	0.79
Patient class prediction 1				
Accuracy	32/41 (0.78)	32/41 (0.78)	32/41 (0.78)	32/41 (0.78)
Specificity	1/9 (0.11)	0/9 (0.00)	0/9 (0.00)	1/9 (0.11)
Sensitivity	31/32 (0.97)	32/32 (1.00)	32/32 (1.00)	32/32 (1.00)
Patient class prediction 2				
Accuracy	36/41 (0.88)	32/41 (0.78)	35/41 (0.85)	36/41 (0.88)
Specificity	6/9 (0.67)	2/9 (0.22)	4/9 (0.44)	4/9 (0.44)
Sensitivity	30/32 (0.94)	30/32 (0.94)	31/32 (0.97)	32/32 (1.00)

Comparing the performance of SVM and CNN in four models, the best performance belongs to Model Corticomedullary using SVM. But considering the fact that CNN gains better performance on a larger dataset, and that there is only a dataset of 11,531 images of 200 SRMs which is not large and balanced enough, it is reasonable to believe that CNN still have room for improvement.

5.4 pRCC/chRCC/ccRCC Classification Results

This section presents all the results related to four models for pRCC/chRCC/ccRCC classification mainly in terms of confusion matrix, accuracy, runtime, and patient predictable diagnosis. Comparison between these models using two different algorithms, SVM and CNN, is discussed at the end of this section.

5.4.1 Model Corticomedullary

The confusion matrix of validation and testing of Model Corticomedullary using SVM is illustrated in Figure 5.7. The confusion matrix of validation is the summation of 10-fold. Generally speaking, the model functioned well on validation subset while performance degraded on the testing subset. For each category, the number of incorrect predictions did not exceed 1/2 of the total number which is a good sign for patient diagnosis in real-world situation.

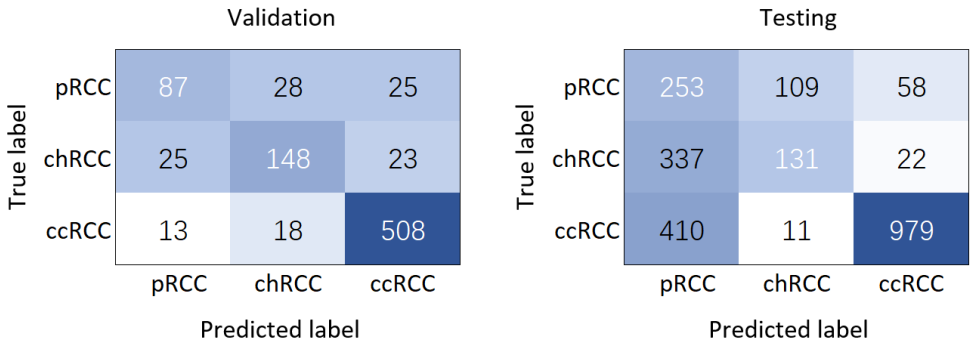

Figure 5.7: Confusion matrix of Model Corticomedullary with SVM

Correspondingly, the confusion matrix of validation and testing of Model Corticomedullary applying CNN is shown in Figure 5.8. The model also worked perfectly on validation subset while exhibited relatively poorer performance on the testing subset. It is evident that for each class the ratio of making correct predictions surpasses 1/2.

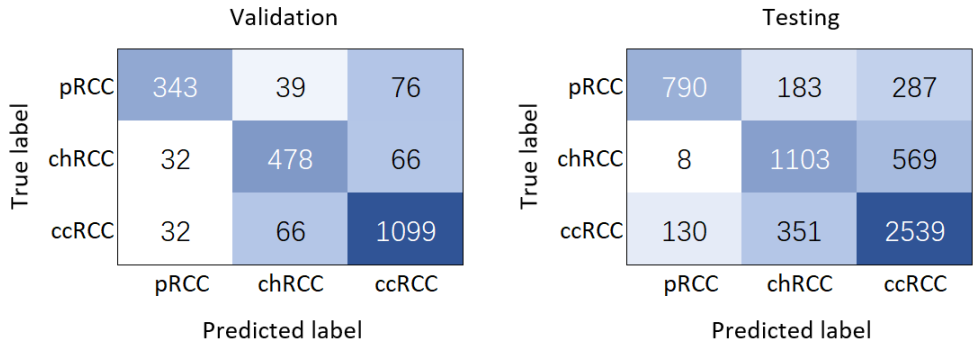

Figure 5.8: Confusion matrix of Model Corticomedullary with CNN

Other evaluation parameters are presented in Table 5.7. Patient class prediction is the same meaning as in fp-AML/RCC differentiation. With the images of Corticomedullary stage, CNN is preferable to SVM, in terms of its performance on validation and testing subsets. It reached an overall accuracy of 0.85 in patient class prediction, in which almost all chRCC and ccRCC cases were predicted correctly and half of pRCC cases were picked out appropriately.

Table 5.7: Performance evaluation of Model Corticomedullary

	SVM	CNN
Validation Accuracy	0.85	0.86
Testing Accuracy	0.59	0.74
Patient class prediction	23/33 (0.70)	28/33 (0.85)
Accuracy for pRCC	5/6 (0.83)	3/6 (0.50)
Accuracy for chRCC	1/7 (0.14)	6/7 (0.86)
Accuracy for ccRCC	17/20 (0.85)	19/20 (0.95)

5.4.2 Model Nephrographic

The confusion matrix of validation and testing of Model Nephrographic using SVM is showed in Figure 5.9. The model performed excellently on validation subset while poorly on the testing subset. Especially for testing, all chRCC cases were picked out but

around 3/4 pRCC cases and 2/5 ccRCC cases were misclassified as chRCC.

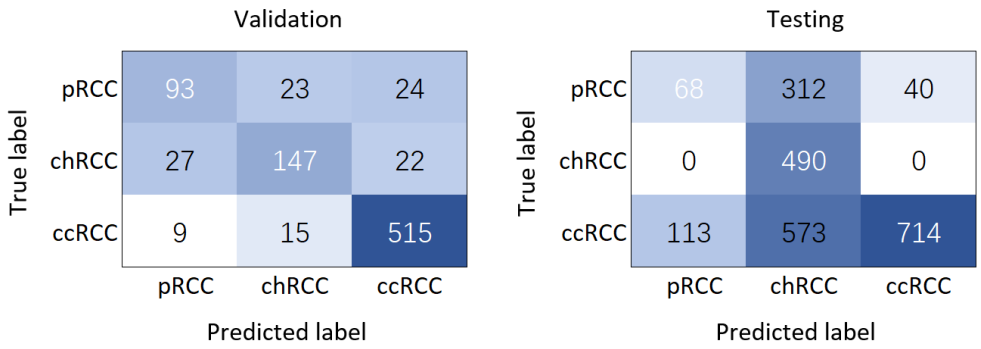

Figure 5.9: Confusion matrix of Model Nephrographic with SVM

In contrast, the confusion matrix of validation and testing of Model Nephrographic applying CNN is shown in Figure 5.10. The model also functioned well on validation subset while exhibited much poorer performance on the testing subset especially for pRCC and chRCC.

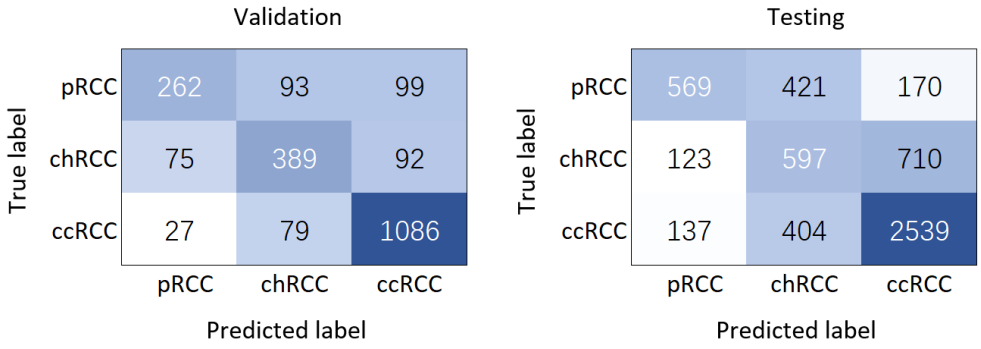

Figure 5.10: Confusion matrix of Model Nephrographic with CNN

Other evaluation parameters are presented in Table 5.8. Patient class prediction is the same meaning as in fp-AML/RCC differentiation. With the images of Nephrographic stage, although SVM performed better than CNN on validation subset, CNN is still a superior choice since it excelled SVM to a great extent on testing subset especially when diagnosing patients.

Table 5.8: Performance evaluation of Model Nephrographic

	SVM	CNN
Validation Accuracy	0.86	0.79
Testing Accuracy	0.55	0.65
Patient class prediction	16/33 (0.49)	25/33 (0.76)
Accuracy for pRCC	0/6 (0.00)	3/6 (0.50)
Accuracy for chRCC	7/7 (1.00)	3/7 (0.43)
Accuracy for ccRCC	9/20 (0.45)	19/20 (0.95)

5.4.3 Model Excretory

The confusion matrix of validation and testing of Model Excretory using SVM is showed in Figure 5.11. On validation set, it had a more balanced performance on three categories while it leaned a bit to chRCC on testing set.

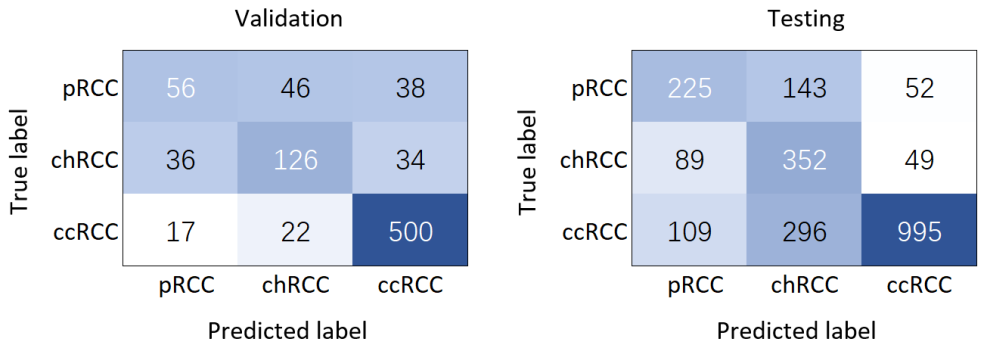

Figure 5.11: Confusion matrix of Model Excretory with SVM

In parallel, the confusion matrix of Model Excretory applying CNN is shown in Figure 5.12. CNN made correct predictions on more than 1/2 of each class on testing set.

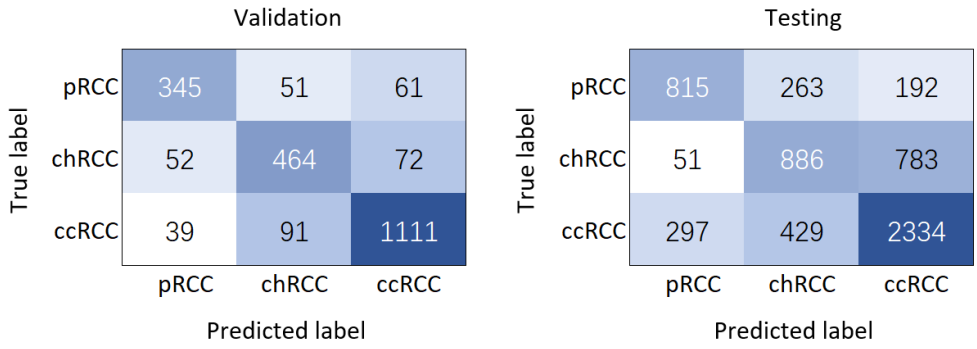

Figure 5.12: Confusion matrix of Model Excretory with CNN

Other evaluation parameters are presented in Table 5.9. SVM and CNN had similar performance with an input of Excretory stage images. The difference lies in that SVM picked out more chRCC cases while CNN detected more ccRCC cases.

Table 5.9: Performance evaluation of Model Excretory

	SVM	CNN
Validation Accuracy	0.78	0.84
Testing Accuracy	0.68	0.67
Patient class prediction	25/33 (0.76)	25/33 (0.76)
Accuracy for pRCC	4/6 (0.67)	4/6 (0.67)
Accuracy for chRCC	6/7 (0.86)	4/7 (0.57)
Accuracy for ccRCC	15/20 (0.75)	17/20 (0.85)

5.4.4 Model Enhanced

The confusion matrix of validation and testing of Model Enhanced using SVM is showed in Figure 5.13. It is obvious that the model performs well on ccRCC detection and poorly on pRCC distinguishment.

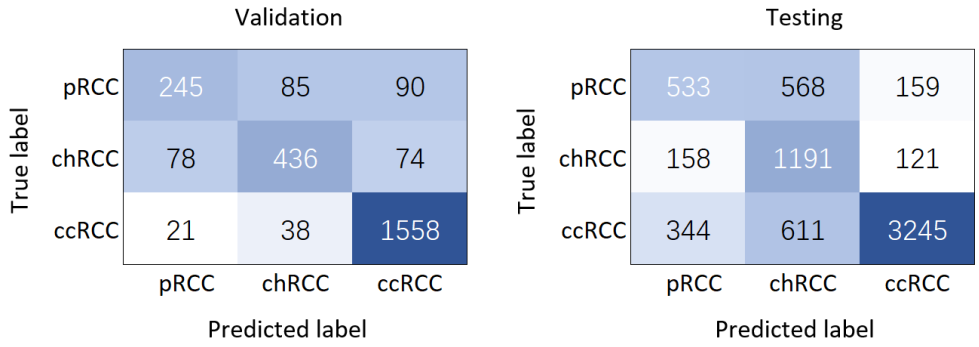

Figure 5.13: Confusion matrix of Model Enhanced with SVM

The confusion matrix of validation and testing of Model Enhanced applying CNN is illustrated in Figure 5.14. It inclined a lot to ccRCC.

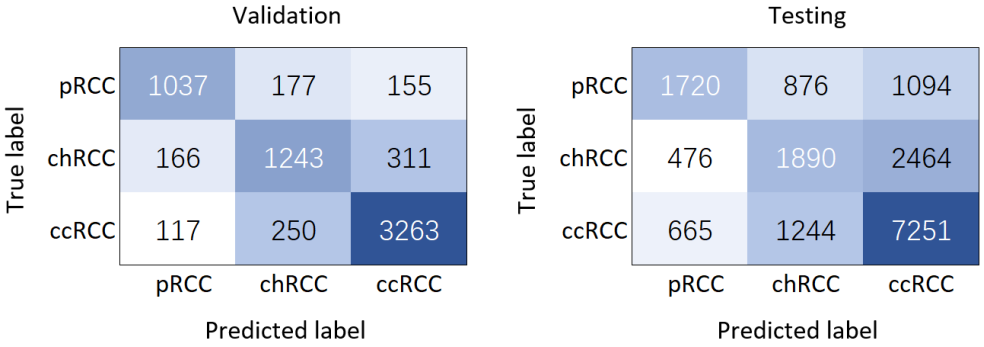

Figure 5.14: Confusion matrix of Model Enhanced with CNN

Table 5.10: Performance evaluation of Model Enhanced

	SVM	CNN
Validation Accuracy	0.85	0.82
Testing Accuracy	0.72	0.61
Patient class prediction	24/33 (0.73)	24/33 (0.73)
Accuracy for pRCC	1/6 (0.17)	3/6 (0.50)
Accuracy for chRCC	6/7 (0.86)	2/7 (0.29)
Accuracy for ccRCC	17/20 (0.85)	19/20 (0.95)

Other evaluation parameters are presented in Table 5.10. SVM and CNN had similar performance with an input of Excretory stage images. The difference lies in that SVM

picked out more chRCC cases while CNN detected more ccRCC cases. SVM excelled CNN on validation and testing accuracy and is the same as CNN in patient class prediction overall accuracy. CNN is more balanced compared to SVM.

5.4.5 Comparison Among Different Models and Algorithms

Comparing four models with two algorithms applied above, CNN with Corticomedullary stage images as input exhibited the best performance which is balanced over three categories. It has attained an overall accuracy of 0.74 on testing subset and 0.85 on patient class prediction. In general, CNN performed better than SVM.

As to runtime, Feature extraction is the most time consuming process in SVM. It took 2.87 h for SVM algorithm to extract features of Corticomedullary stage images. The 10-fold training, validation, and testing process took 6.75 s in total. It took less than 0.01 s to test on each patient, which is acceptable. It took around 15.81 h for the 10-fold training and cross validation of Model Corticomedullary by using CNN. The testing process took 14 s for all patients and less than 0.5 s for each patient diagnosis. They are both rapid enough for real-world practices.

5.5 Summary

In this chapter the results of four models, Model Corticomedullary, Model Nephrographic, Model Excretory, and Model Enhanced with SVM and CNN algorithms, are elaborated and analyzed. Images of Corticomedullary stage has been revealed to be qualified for SRM classification. SVM is strong at fp-AML/RCC classification with an accuracy of 0.93 and a sensitivity of 0.96 on patient diagnosis. CNN has its advantage in pRCC/chRCC/ccRCC differentiation reaching an accuracy of 0.85 on patient tumor type prediction. They both showed great potential in that SVM exceeded what has been achieved by former studies in various evaluation methods and CNN opened up a new direction in SRM CT image processing.

Chapter 6

Conclusion and Discussion

6.1 Conclusion

Eight models with two algorithms, SVM and CNN, and four varied input, corticomedullary, nephrographic, excretory, and enhanced stage CT images of SRMs, were proposed and developed for SRM classification in this study. Model Corticomedullary using SVM is qualified for fp-AML/RCC classification with an accuracy of 0.93 and a sensitivity of 0.97 on patient diagnosis. Its performance has surpassed that of previous studies. Model Corticomedullary using CNN is capable of pRCC/chRCC/ccRCC categorization and achieved an accuracy of 0.85 on patient class prediction. It is a breakthrough since it is the first successful application of deep learning networks in kidney tumor image processing. In addition, the results of both models are balanced over different classes. They together offer a comprehensive solution to SRM classification. Given these findings, the two algorithms can be a preliminary step in the advancement of a decision model which may serve an important role in guiding appropriate management and treatment in clinical practice.

6.2 Evaluation of My Work

The main outcome of my work is two models respectively with SVM and CNN applied for fp-AML/RCC and pRCC/chRCC/ccRCC classification. No previous researches has been done on this topic of comprehensive SRM classification. Recently, several publications involved utilizing conventional machine learning aided approaches on similar clinical dilemma by extracting handcrafted diagnostic features from CT images. They reported moderate to good diagnostic performance with an accuracy of 0.72-0.94 but most of them only had a small amount of efficient raw image data and only applied hold-out cross validation [30][32][38][39]. This study has implemented 10-fold cross validation and achieved a high accuracy of 0.93 and satisfying sensitivity of 0.97 on real case diagnosis. As deep learning hasn't been exploited in this field before, this study collaborated CNN and tumor CT image processing to tackle the conventional time and resource consuming problem and has proved to be promising as it reached an accuracy of 0.85, balanced over three categories. These two state-of-art models can be seen as important improvements on utilizing algorithms in SRM differentiation.

6.3 Future Work

There are still ways to optimize models and improve this study.

Firstly, the dataset used in this research is not in very large scale and it is quite imbalanced over benign and malignant cases. This is due to the fact that it is difficult for one hospital to collect enough effective images and that fp-AML accounts for only a small fraction of all SRMs. An imbalanced dataset can lead to the decision made by deep learning algorithm lean to the class with larger number of input images. It is very likely that a larger and more balanced dataset will improve the performance of deep learning algorithm and this can be realized by promoting cooperation among more hospitals and establishing a large scale multi-center study, which is required to fully exploit the generalization capacity of deep learning algorithm.

Secondly, the SVM can be optimized by feature selection approaches such as genetic algorithm. Other multi-label classification methods are also options since one vs the rest

strategy did not show satisfactory results. CNN can be improved by adjusting multiple model parameters and applying regularization functions and earlystopping strategy for the prevention of overfitting.

Thirdly, tumors in this study were contoured manually by a radiologist, which is time consuming as segmentation of one patient took up to 25 minutes. It highlights the need for automatic tumor segmentation.

References

[1] Torre L A, Bray F, Siegel R L, et al. Global cancer statistics, 2012[J]. CA: A Cancer Journal for Clinicians, 2015, 65(2): 87-108.

[2] Chen W, Zheng R, Baade P D, et al. Cancer statistics in China, 2015[J]. CA: A Cancer Journal for Clinicians, 2016, 66(2): 115-132.

[3] Richard P O, Jewett M A S, Bhatt J R, et al. Renal tumor biopsy for small renal masses: a single-center 13-year experience[J]. European Urology, 2015, 68(6): 1007-1013.

[4] Low G, Guan H, Fu W, et al. Review of renal cell carcinoma and its common subtypes in radiology[J]. World Journal of Radiology, 2016, 8(5):484-500.

[5] Ljungberg B, Bensalah K, Canfield S, et al. EAU guidelines on renal cell carcinoma: 2014 update[J]. European Urology, 2015, 67(5): 913-924.

[6] Lei Z, Li X S, Zhou L Q. Natural History of Small Renal Masses[J]. Chinese Medical Journal, 2015, 128(9):1232-1237.

[7] Nelson C P, Sanda M G. Contemporary diagnosis and management of renal angiomyolipoma[J]. The Journal of Urology, 2002, 168(4):1315-1325.

[8] Lane B R, Aydin H, Danforth T L, et al. Clinical correlates of renal angiomyolipoma subtypes in 209 patients: classic, fat poor, tuberous sclerosis associated and epithelioid[J]. The Journal of Urology, 2008, 180(3): 836-843.

[9] Hou W, Xiao H, Liu G, et al. Angiomyolipoma being surgically excised for presumed kidney carcinoma[J]. International Urology and Nephrology, 2015, 47(7): 1037-1043.

[10] Turajlic, Samra, Larkin, et al. SnapShot: Renal Cell Carcinoma[J]. Cell, 2015, 163(6):1556.

[11] Ramos J, Kava B, Leveillee R, et al. Enhancement characteristics on 4 phase dedicated renal protocol CT helps to distinguish benign from malignant tumors [J]. The Journal of Urology, 2011, 185(4): e282-e283.

[12] Jinzaki M, Silverman S G, Akita H, et al. Renal angiomyolipoma: a radiological classification and update on recent developments in diagnosis and management[J]. Abdominal Imaging, 2014, 39(3): 588-604.

[13] Campbell S C, Novick A C, Belldegrun A, et al. Guideline for management of the clinical T1 renal mass[J]. The Journal of Urology, 2009, 182(4): 1271-1279.

[14] Hélénon O, Eiss D, Debrito P, et al. How to characterise a solid renal mass: a new classification proposal for a simplified approach[J]. Diagnostic and Interventional Imaging, 2012, 93(4): 232-245.

[15] Pierorazio P M, Hyams E S, Tsai S, et al. Multiphasic enhancement patterns of small renal masses (≤ 4 cm) on preoperative computed tomography: utility for distinguishing subtypes of renal cell carcinoma, angiomyolipoma, and oncocytoma[J]. Urology, 2013, 81(6): 1265-1272.

[16] Gillies R J, Kinahan P E, Hricak H. Radiomics: images are more than pictures, they are data[J]. Radiology, 2015, 278(2): 563-577.

[17] Simpfendorfer C, Herts B R, Motta-Ramirez G A, et al. Angiomyolipoma with minimal fat on MDCT: can counts of negative-attenuation pixels aid diagnosis?[J]. American Journal of Roentgenology, 2009, 192(2): 438-443.

[18] Simpson E, Patel U. Diagnosis of angiomyolipoma using computed tomography—region of interest≤− 10 HU or

4 adjacent pixels≤− 10 HU are recommended as the diagnostic thresholds[J]. Clinical Radiology, 2006, 61(5): 410-416.

[19] Catalano O A, Samir A E, Sahani D V, et al. Pixel distribution analysis: can it be used to distinguish clear cell carcinomas from angiomyolipomas with minimal fat?[J]. Radiology, 2008, 247(3): 738-746.

[20] Xie P, Yang Z, Yuan Z. Lipid-poor renal angiomyolipoma: Differentiation from clear cell renal cell carcinoma using wash-in and washout characteristics on contrast-enhanced computed tomography[J]. Oncology Letters, 2016, 11(3): 2327-2331.

[21] Kim M, Lee J B, Cho G, et al. MDCT-based scoring system for differentiating angiomyolipoma with minimal fat from renal cell carcinoma[J]. Acta Radiologica, 2013, 54(10): 1201-1209.

[22] Kim J K, Park S Y, Shon J H, et al. Angiomyolipoma with minimal fat: differentiation from renal cell carcinoma at biphasic helical CT[J]. Radiology, 2004, 230(3): 677-684.

[23] Giger M L, Chan H P, Boone J. Anniversary paper: History and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM[J]. Medical Physics, 2008, 35(12): 5799-5820.

[24] Gurusamy R, Subramaniam V. A machine learning approach for MRI brain tumor classification[J]. Computers, Materials & Continua, 2017, 53 (2): 91-109.

[25] Tu S J, Wang C W, Pan K T, et al. Localized thin-section CT with radiomics feature extraction and machine learning to classify early-detected pulmonary nodules from lung cancer screening[J]. Physics in Medicine and Biology, 2018, 63(6): 065005.

[26] Wang H, Zhou Z, Li Y, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 F-FDG PET/CT images[J]. EJNMMI Research, 2017, 7(1): 1-11.

[27] Shah A. Prediction of malignant and benign tumor using machine learning[J]. International Journal of Computer Applications, 2016, 135(5): 19-23.

[28] Garapati S S, Hadjiiski L, Cha K H, et al. Urinary bladder cancer staging in CT urography using machine learning[J]. Medical Physics, 2017, 44(11): 5814-5823.

[29] Woo S, Cho J Y, Kim S H, et al. Angiomyolipoma with minimal fat and non-clear cell renal cell carcinoma: differentiation on MDCT using classification and regression tree analysis-based algorithm[J]. Acta Radiologica, 2014, 55(10): 1258-1269.

[30] Lee H S, Hong H, Jung D C, et al. Differentiation of fat-poor angiomyolipoma from clear cell renal cell carcinoma in contrast-enhanced MDCT images using quantitative feature classification[J]. Medical Physics, 2017, 44(7): 3604-3614.

[31] Vieira S, Pinaya W H L, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications[J]. Neuroscience & Biobehavioral Reviews, 2017, 74: 58-75.

[32] Yan L, Liu Z, Wang G, et al. Angiomyolipoma with minimal fat: differentiation from clear cell renal cell carcinoma and papillary renal cell carcinoma by texture analysis on CT images[J]. Academic Radiology, 2015, 22(9): 1115-1121.

[33] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017, 42: 60-88.

[34] Yasaka K, Akai H, Abe O, et al. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study[J]. Radiology, 2017, 286(3): 887-896.

[35] Jones N. The learning machines[J]. Nature, 2014, 505(7482): 146.

[36] Kochhar R, Brown R K, Wong C O, et al. Pictorial Essay: Role of FDG PET/CT in imaging of renal lesions[J]. Journal of Medical Imaging and Radiation Oncology, 2010, 54(4): 347-357.

[37] Blazic I M, Lilic G B, Gajic M M. Quantitative assessment of rectal cancer response to neoadjuvant combined chemotherapy and radiation therapy [J]. Radiology, 2017, 282(2): 614.

[38] Hodgdon T, McInnes M D F, Schieda N, et al. Can quantitative CT texture analysis be used to differentiate fat-poor renal angiomyolipoma from renal cell carcinoma on unenhanced CT images?[J]. Radiology, 2015, 276(3): 787-796.

[39] Takahashi N, Leng S, Kitajima K, et al. Small (< 4 cm) renal masses: differentiation of angiomyolipoma without visible fat from renal cell carcinoma using unenhanced and contrast-enhanced CT[J]. American Journal of Roentgenology, 2015, 205(6): 1194-1202.

Acknowledgement

First of all, I would like to express my sincere appreciation to all those who helped me to complete this thesis. A special gratitude I give to my supervisors, Prof. Jussi Toivonen in UTU and Prof. Guohui Zhou in FDU for their patient guidance, friendly support, and strict requirements for this thesis.

Thanks to PhD Xiao Li who gave me access to adequate dataset and essential medical and radiological knowledge.

I also owe a deeply gratitude to all the professors in my lab who equipped me with inspiring ideas and professional opinions. Thanks to all the lecturers and classmates both in FDU and UTU for their efforts that let me experienced a memorable life during my graduate career.

Finally, thanks to my beloved family and my boyfriend Luning. Without their support I cannot finish this thesis.