**CHAPTER 2**

**BACKGROUND AND LITERATURE REVIEW**

**2.1 Introduction**

A review of the background and previous work with reference to the

research illustrated in thesis is presented in this chapter. The content of this chapter is

organised as follows Section 2.2 Medical Image, a brief history of medical image

technology and the implementation of medical image in human debases diagnose,

typically Knee-OA is presented. Section 2.3 Image Processing, the basic idea of image

processing technology is described. There are three image processing techniques: (i)

Image Segmentation, (ii) Image Enhancement and (iii) Feature Extraction/Image

Representation. The brief detail of each technique is presented in Sub-section 2.3.1,

Sub-section 2.3.2 and Sub-section 2.3.3, respectively. The summary information of

image classification is suggested in Section 2.4. The detail of the section is described

as follows: (i) the fundamental idea of data mining (ii) the detail of image classification

and (iii) the classification learning methods are presented Sub-section 2.4.1 to 2.4.3,

respectively. The feature selection techniques which apply to select the feature vector

are described in Section 2.5. Section 2.6 points out of the brief information of

Convolutional Neural Network (CNN) applying in medical imaging. The evaluation

measurement methods which is used to evaluate the classification result is discussed in

Section 2.7, in this section there are six measure parameters including: (i) Accuracy

(AC), (ii) Sensitivity (SN), (iii) Specificity (SP), (iv) Precision (PR), (v) F- Measure

(FM), and (iv) Area under the ROC Curve (AUC). In case of AC, SN, SP, PR, and FM

can be calculated from Confusion Matrix which is presents in Sub-section 2.7.1. While

AUC can calculates from the Receiver operating characteristic (ROC) curve which is

10

presented in Sub-section 2.7.2. The comparison of some related works are presented in

Section 2.8. Finally, the summary of this chapter is described in section 2.9.

**2.2 Introduction**

Medical image has spread widely and has been an interdisciplinary to

many research fields, for example, computer science, computer vision, engineering,

statistic, biology, and medicine. The technology in computer-aided diagnose processing

has become an important for clinical healthcare. The medical image is used to diagnose

inside human structure such as bone, brain and renal system. There are various types of

medical image with respect to image capturing mechanisms including (i) X-ray

(radiography), (ii) Computed Tomography (CT), (iii) Magnetic Resonance Imaging

(MRI), (iv) Ultra-sound, (v) Positron Emission Testing (PET) and (vi) Single Photon

Emission Computed Tomography (SPECT). With respect to the work presented in this

research, to act as a focus for the work X-ray image is considered. However, there is

three medical image capture mechanisms are presented in this section including (i) X-

ray, (ii) CT and (iii) MRI.

2.2.1 X-ray Image

German physicist Wilhem Rontgen discovered X-ray image in 1895 [6].

The first medical image of X-ray image was taken from his wife’s left hand. Wilhelm

Rontgen took this radiograph of his wife's left hand on 22 December 1895, shortly after

his discovery of X-rays. The image of Wilhelm’s wife’s hand is shown in Figure 2.1.

11

Figure 2.1 The image of Wilhelm wife left hand [6].

X-ray is a form of electromagnetic radiation which can pass through the

human body with producing the internal structure image, the resulting of X-ray is an

image, and this image is called a radiographic, also known as “X-Ray” or “plain film”.

Figure 2.2 An example of X-ray image with different radiographic density [7].

There are five categories depending on radiographic densities in X-ray

image. The Figure 2.2 illustrates the difference of X-ray radiographic density indicating

by the number including (i) number 1 is air (the darkest area), (ii) number 2 presents

12

fat (the dark grey), (iii) number 3 is the soft tissue (grey), (iv) number 4 illustrates bone

(bright grey) and (v) number 5 is the contrast material [7].

X-ray images have been used to diagnose Knee-OA which presented in

many works [8,9,10,11,12,13,14,15]. There are two main objectives for Knee-OA

research: (i) X-ray image was used to detect OA and non OA [9,11], while the second

one is more advance than the first one because the X-ray image has been used to detect

the OA stages as reported in [10,12,13,14,215].

On the purpose of the detection of the OA stage, the Kellgren-Lawrence

(KL) system [16,17] is considered. The KL standard is a five stages indicating using

stage 0 to stage 4. ‘Stage 0’ represents no OA symptom while ‘Stage 4’ represents the

most serve radiographic disease. The summary of the OA stage is shown in Table 2.1.

There are five stage of OA has shown in the table which rank from zero to fourth stage.

The stage 0 refers to the normal case, within this stage the sign of OA cannot detect

because there is no sign of OA has happened. The second stage of OA is the doubtful

stage sometime called stage 1 of knee OA; in this stage there is a little sign of OA has

appeared. In other words, in the stage 1 of OA there is a sign to show very minor bone

spur growth. The next stage is the stage 2 are considered as the minimal stage, the knee

joints in this stage will reveal greater bone spur growth, but the cartilage is usually still

at a healthy size, i.e. the space between the bones is normal, and the bones are not

rubbing or scraping one another. At this stage, synovial fluid is also typically still

present at sufficient levels for normal joint motion. However, this is the stage where

people may first begin experiencing symptoms; pain after a long day of walking or

running, greater stiffness in the joint when it’s not used for several hours or tenderness

when kneeling or bending. Stage 3 of OA is considered as “moderate” stage. In this

stage, the cartilage between bones shows obvious damage and the space between the

bones begins to narrow. People with stage 3 OA of the knee are likely to experience

frequent pain when walking, running, bending, or kneeling. They also may experience

joint stiffness after sitting for long periods of time or when waking up in the morning.

Joint swelling may be presented after extended periods of motion. Stage 4 OA is

considered as “severe.” People in stage 4 OA of the knee experience great pain and

discomfort when they walk or move the joint. This is because the joint space between

bones is dramatically reduced; the cartilage is almost completely gone, leaving the joint

13

stiff and possibly immobile. The synovial fluid is decreased dramatically, and it no

longer helps reduce the friction among the moving parts of a joint.

Table 2.1 The summary of the knee OA stages

Knee OA

Stage

*Stage 0*

(Normal)

*Stage 1*

(Doubtful)

*Stage 2*

(Minimal)

*Stage 3*

(Moderate)

*Stage 4*

(Severe)

Properties No sign of

OA

Show very

minor bone

spur growth

Reveal

greater

bone spur

growth

Cartilage

was damage

and

narrowing

of joint

space

Great pain

and

discomfort

when

walking

Figure 2.3 shows the X-ray image of each OA levels Figure 2.33(a)

presents stage 0 OA or normal, Figure 3 (b) illustrates stage 1 OA (doubtful), Figure

3(c) shows stage 2 OA the minimal level and Figure 3 (d) presents stage 3 moderate

level.

Figure 2.3 X-ray image of four different KL grade [14]

14

The work of OA stage detection using X-ray offers the benefit over the

detection the OA and non-OA because knowing the level of OA can make orthopaedic

doctor easy to prevent and slow down the OA speed to the OA patient. The most

important is the cost because the price of X-ray image is unexpansive compare to the

others.

2.2.2 Computed Tomography

Computed Tomography (CT) is one of the well-known medical imagery

for disease diagnosis. A CT scan uses a collection of X-ray images taken from different

angles to produce cross-sectional image using the computer processing. Thus the

structures inside of the body can be presented without cutting. CT provides the clear

image so the medical doctor can be used to analyse or diagnose the disease.

2.2.3 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) uses the magnetic properties of

spinning hydrogen atoms to make the internal structure image. It makes the powerful

image of human bone structure and create a very clear image. With respect to OA

detection, the MRI images were also applied as presented in [18,19,20,21]. The works

show the benefit of OA analysis with short time consuming and produce a good result.

The example of MRI image of human knee is shown in Figure 2.4.

15

Figure 2.4 Human knee image using MRI [18]

To summarise, OA detection is the challenge task in medical and computer science. The image processing is applied to OA detection using medical image is the best solution as well as to analyse the OA stage. With the respect to this research to act as a focus in Thailand, so the medical X-ray image is considered with low cost of processing. With X-ray image apply with image processing to diagnose OA stage, the image processing technique is presented in the next section.

**2.3 Image Processing**

The fundamental concepts of image processing is discussed in this

section. Image processing is a challenge area in computer science. The aims of this

method is to analyse and manipulate the digital image. In addition image processing

has been applied to improve the image quality. There are main three key words in image

processing that can be listed as follows:

(i) Image processing (input as image and output as image)

16

(ii) Image analysis (input as image and output as the measurement)

(iii) Image understanding (input as image and output as the high-level

description)

An image is a set of array, a matrix, a square of pixel which arrange in

columns and rows. In 1920s, the first image processing was discovered and it was called

Bartlane cable picture transmission system. Nowadays, image processing has been

applied in various applications such as education, engineering and medical. To get the

deep meaning of image processing, the definition of an image should be known clearly

first. An image is considered of two real variable, for example I(x,y), with I as the

amplitude of the image at the real coordinate position (x,y). Sub-image is the sub set

of image, sometime it considers as the region of interest (ROI). The smallest element

of image is pixel. Image processing generally separate into two main categories: (i)

digital image processing and (ii) medical image processing [91].

In general the medical image processing provides the image in white and

black colour, or greyscale image in technical word. An 8-bits greyscale image has

which picture element has an assigned intensity that ranges from 0 to 255. Figure 2.5

shows each pixel of greyscale image with the value from 0(black) to 255 (white)

because 8 bit equal to 256 greyscales.

Figure 2.5 Each pixel has a value from 0 (black) to 255 (white) [22]

There are three image processing processes including: (i) image

segmentation, (ii) image enhancement and (iii) image representation/image feature

extraction. The detail of each process is discussed in the Sub-section 2.3.1 to 2.3.3

respectively.

17

2.3.1 Image Segmentation

Image segmentation is one of the most important processes of the image

processing. The main objective of image segmentation is to divide the given image in

to salient image regions, meaningful region, and homogeneous with specific cluster

pixel. In addition, the image segmentation is applied to various application domains

includes: object recognition, occlusion boundary estimation within motion or stereo

systems, image compression, image editing, or image database look-up. A good

segmentation of an image can be defined by: (i) pixel in the same region or categories

has similar or equal pixel colour/intensity and (ii) the neighbouring pixel in different

categories has different values or dissimilar values.

Although image segmentation is one of the main process in image processing, but it has

many algorithms to apply with based on the different kind of work. Some of existing

image segmentation methods are presented here: (i) threshold-based segmentation, (ii)

region-based segmentation and (iii) edge-based segmentation. The brief detail for each

method is discussed as follow:

**Threshold segmentation**

Threshold-based segmentation is one the existing methods for image

segmentation. The separation of bright and dark region is the main objective of

threshold segmentation. In general, thresholding is applied to create the binary image.

For all pixel of greyscale image is calculated by comparing each pixel with threshold

(T). If the intensity values is below than threshold the pixel which make from grey-

level ones by turning all pixels that below threshold become zero (0) and if pixel is

above threshold the pixel become one (1). For example: if I(x,y) is a threshold version

of i(x,y) at some global threshold T, as Equation 1.1 belows:

I(x,y)

={

1 0 if otherwise

i(x,y) ≥ T

1.1

Threshold-based segmentation has a drawback which is only the pixel

value is used to consider on other hand no relationship between the pixels is examined.

Therefore, there is no any guarantee that the pixels identified by the thresholding

process are contiguous.

18

Figure 2.6 The human brain scan using Otsu method [20].

There are many applications using threshold mechanisms were proposed

for medical image segmentation. In [20] the implementation of Otsu method to improve

medical image segmentation. Otsu method is a method to reduce the greyscale to a

binary image, the result shown that the method performs better than other thresholding

methods with a good binary images. Otsu threshold method is a very good and efficient

method to threshold the greyscale image. However, the drawback of this method is high

complexity of the computation. Figure 2.6 (a) and Figure 2.6 (c) shows the original of

human brain scan, while Figure 2.6 (b) and Figure 2.6 (d) presents the corresponding

segmented image using Otsu method.

Local adaptive thresholding also is applied in [21]. The work reported

the implementation of local adaptive threshold technique to removes background using

local mean and standard deviation. The comparative study between Niblack and

Sauvola local thresholding had been evaluated. The results shown that the Niblack local

threshold algorithm reduced the background noise better than Sauvola local threshold

algorithm. Figure 2.7 shows the difference of five sample sub-images between two

different methods: Niblack technique and Sauvola technique.

19

Images Using Niblack Technique Using Sauvola Technique Image 1

Image 2

Image 3

Image 4

Image 5

Figure2. 7 Example of The comparison of five medical images by using Niblack and

Sauvola Technique [21].

**Region based segmentation**

Region-based segmentation is one of image segmentation techniques.

The main objective is to find coherent regions in the image. The coherent region

20

contains pixels that has same or some similar property. The transitivity of the similarity

relationship in the image is considered. Region-based segmentation offers the

advantages on (i) fast processing and (ii) more efficient than edge-based and threshold-

based method. However the region-based also has the drawback concerning it probably

grow further away from global pixel because it drifts as one. There are many existing

region-based segmentation algorithms for example watershed segmentation, flooding-

based watershed segmentation, marker-controlled watershed segmentation and inter-

pixel watershed segmentation. The implementation of watershed segmentation in the

context of medical image with texture-based region merging was proposed in [23]. The

work in [33] reported the texture-based region merging was applied to MRI images in

order to improve the segmentation efficiency. The comparison of images using

watershed segmentation with/without text-based region merging and with texture-based

region-based region merging is illustrated in Figure 2.8.

Figure 2.8 The comparison of image using watershed segmentation with/without

texture-based [23]

21

**Edge based segmentation**

In edge-based segmentation mechanism focuses on the edge in an

image. The edge corresponds to singularities in the images. The edge in image generally

represents as the shape of objects in the scene. The purpose of edge-based segmentation

is to extract the edge or line in the image with good orientation. There are some existing

edge operators for example Gradient operator, Prewitt operator, Sobel operator,

Compass operators and Laplacian operator. Figure 2.9 shows the Prewitt and Sobel

operator.

Figure 2.9 Prewitt and Sobel operator for edge detection

With the respect to the edge detection in the context of medical image

segmentation, it can be found in [24]. The edge detection was applied to the brain scan

images to detect the tumour regions based on the gradient magnitude information. The

example image after applied the edge detection is shown in Figure 2.10.

Figure 10 The examle of brain scan image using edge detection [24]

22

From Figure 2.10 (a) shows the original brain scan while Figure 2.10 (b)

illustrates the brain scan image after the edge detection was applied.

2.3.2 Image Enhancement

Image enhancement is one of the most important and complex

techniques in image processing. The fundamental idea of image enhancement

mechanism is to improve visual appearance of an image, or to present a better transform

representation of the image. The main challenge in image enhancement is quantifying

the criterion for enhancement.

There are some varies types of image such as personal images, medical

images, satellite images and aerial images. The photographers/users/viewers probably

suffer from some image quality problems such as poor contrast and noise. Therefore

image enhancement plays the main role to enhance the contrast and remove the noise

to increase image quality. During image enhancement process an input is an original

image and the output also an image with better than the input image by changing the

pixel intensity. Furthermore, image enhancement works as the important roles in many

kinds of image for example hyper spectral image processing, remote sensing, high

definition television (HDTV), industrial X-ray image processing, microscopic imaging,

other image/video processing applications and medical image processing [25].

In addition image enhancement technique can be used to increase

dynamic range of the chosen features in of an image such as point operation, spatial

operation, transform operation and pseudo colouring. There are some existing methods

of image enhancement for example: filtering with morphological operators, histogram

equalization, noise remove using a Wiener filter, linear contrast adjustment, median

filtering, unsharp mask filtering, contrast-limited adaptive histogram equalization

(CLAHE) and decorrelation stretch.

23

Figure 2.11 The example of X-ray image of renal system using morphological

transform [26]

In the context of medical image, image enhancement also applied in

many applications. Morphological transform was applied to medical images [26]. The

work reported on using a disk shaped mask is used in Top-Hat and Bottom-Hat

transform in order to remove noise and increase the contrast quantity in medical images.

Figure 2.11 shows the morphological transform was used in X-ray of renal system.

Figure 2.11 (a) presents the original image. Figure 2.11 (b) illustrates optimum

enhanced image. While Figure 2.11 (c) and Figure 2.11 (d) shows the transformed

image with different radius using Carrier-to-noise ratio (CNR) and Peak signal-to-noise

ratio (PSNR), respectively.

2.3.3 Feature Extraction or Image Representation

The fundamental information of feature extraction or image

representation is presented in this sub-section. In image processing, feature extraction

is the main rule for measuring the data and builds derived values (features) in order to

find a good data representation. The objective of feature extraction in digital image is

to reducing the amount of resources which required to describe a large set of data.

Feature extraction is directed relation to dimensional reduction; dimensional reduction

is the process of reducing the number of random variables under consideration [27].

With the respect to dimensional reduction there are some existing techniques including:

Independent component analysis, isomap, principal component analysis (PCA), kernel

principle component analysis, latent semantic analysis, partial least squares, multifactor

24

dimensionality reduction, nonlinear dimensionality reduction, multilinear principal

component analysis, multilinear subspace learning, Semidefinite embedding,

autoencoder and deep feature synthesis.

However in the context of image representation or feature extraction in

digital images can be done using various using image properties. There are three main

image properties including: (i) colour, (ii) texture and (iii) shape. The detail of each

image feature extraction technique is described as follows:

**Colour analysis**

Colour is simplest features in digital image. It makes an image is more

colourful and more interesting to human view [92]. The colour analysis for image

representation is a technique that can use widely with RGB image. RGB image is one

type of digital image which contain RGB color system; RGB system refers to the

system that representation in Red, Green and Blue through computer display view. The

level of red, green and blue can range from 0 to 100 percent of full intensity. For each

level of red, green and blue is represented by the range of decimal numbers from 0 to

255 (256 levels for each colour), or to be a binary is 00000000 or 11111111. RGB also

know as true colour image. With respect to digital image which store as an m-by-n-by-

3 (m x n x 3) data array that defines red, green, and blue colour components for each

individual pixel. The purpose of colour image analysis is to define or calculate the

percent of each colour has contained in an image.

**Texture analysis**

The texture analysis concerns with description of characteristic image

properties by textural features. Texture analysis [28,29] is an important approach for

describing the region. Texture can be divided into three categories: (i) statistical

methods (a collection of statistic that used to described of texture), (ii) structural

methods (texture is viewed as consisting of many texels arranged according to some

placement rules ) and (iii) model based methods (used probability model to modelled

the image textures). A number of existing texture analysis models has been proposed

including autoregressive model, Gausian Markov random field, Gibbs random fields,

Wold model, wavelet model, multichannel Gabor model and steerable pyramid.

25

Texture analysis is also implemented in the context of medical image.

The example is found in [11] the long rage, standard deviation and greyscale entropy

texture analysis techniques were applied to X-ray images to detect OA and non-OA,

the Haralick feature was calculated in order to measure the texture image in term of

contrast, correlation, sum of square, sum of average and homogeneity.

In [19], Frequency Filtering texture analysis was applied using Fourier

transform to calculate filter data as the Equation 2.2

VD = F \* VD 2.2

where

VD is the filter data from the convolution theorem.

F is the filter in the spatial domain.

The examples various filters are illustrates in Figure 12 including: (i)

Low pass filter (Figure 2.12 (a)), (ii) High pass filter (Figure 2.12 (b)) and (iii) Band

pass filter (Figure 2.12 (c)). The results after applied given filters to human knee MRI

image are presented in Figure 2.12 (d), Figure 2.12 (e) and Figure 2.12 (f), respectively.

Figure 2.12 The example of various filters applied to MRI of human

26

**Texture analysis**

Shape analysis is directed to study the edge of shape feature in image.

Shape refers to all the geometrical information that remains when location, scale and

rotational effects are filtered out from an object [30]. Shape is any connection of point

in the image. The example of the existing shape analysis mechanisms including: point

distribution, active shape, active appearance models, Fourier Snakes, active contours

and parametrically-deformable models. The shape analysis was applied in the context

of medical image in many literatures. With respect to the work in this research is

directed to Knee OA. There are some examples found in [9, 18, 31].

In [9] the shape analysis was used to calculate the thickness between the

bone and joint space in knee. Thickness was calculated and found that in the range of

1.69 mm to 2.55 mm was non-OA. In contrast if the thickness is less than 1.69 mm then

the symptom of OA was possible. Figure 13 shows the cropping image and applies with

binary operation to make edge detection in order to detect the boundary between the

bone and joint space.

Figure 2.13 Cropping the image and apply in binary operation [9]

With the respect to the work presented in [31], the shape analysis was

applied to calculate the area of cartilage. The initial step was the pixel-based

segmentation is used to segment the cartilage. The texture filtering was then applied to

calculate the area of cartilage by the number of pixel as presented in Figure 2.14.

27

Figure 2.14 MRI image of knee after the texture filtering [31]

In summary, the image segementation, image enhancement and image

representation are the three main process in digital image processing. With the respect

to the work presented in this thesis the image segementation, enhancement and

representation were applied as the pre-procesing process. The next section is 2.4 image

classification is discussed.

**2.4 Image Classification**

The basic information of image classification is presented in this section.

With the respect to this work in this thesis the image classification mechanism is

commenced. The fundamental idea of data mining is presented in Sub-section 2.4.1.

Sub-section 2.4.2 suggests the detail of image classification and finally, Sub-section

2.4.3 presents the brief detail of classification learning methods.

28

2.4.1 Data Mining

Data mining is the process of discovering and searching insightful,

interesting, useful, as well as descriptive, understandable and predictive models from

large-scale data [32]. Data mining considered as the non-trivial of novel, implicit and

actionable knowledge from large datasets. The alternative terms refer to knowledge

mining from database, knowledge extraction, data analysis and data archaeology.

In addition, data mining is considered as the core element in the

knowledge discovery in database (KDD) process .The phase knowledge discovery in

database was suggested in the first KDD workshop in 1989. The KDD processes

including data selection, data cleaning, data transformation, pattern searching (data

mining), patterns/finding evaluation, data visualization and knowledge representing.

In [33] KDD consists of five processes: (i) Selection, (ii) Preprocessing,

(iii) Transformation, (iv) Data mining and (v) Interpretation and evaluation as shown

in Figure 2.15. From the Figure 2.15 can be observed that the KDD is a five-process

mechanism. During selection process a large collection of data is selected. Once data

selection is completed, the transformation process is then concerned in order to prepare

appropriate data format to be ready for the data mining process [33]. During the data

mining process, the data mining algorithms or techniques are implemented. When the

pattern or hidden information is delivered the interpretation and evaluation process is

applied to get the accurate knowledge.

Figure 2.15 KDD process for producing the knowledge.

29

Data mining can be divided into two categories: (i) predictive data

mining and (ii) descriptive data mining. Predictive data mining concerns with the

prediction of behaviour based on historical data. In contrast descriptive data mining is

directed at the discovery of patterns in existing data in order to use as the guideline for

making the decision in the future.

In [34] presented the tasks of data mining. In general data mining task

is divided into three main categories: (i) Classification, (ii) Association and (iii)

Clustering.

Classification is the process of finding a model that describes the data

classes or concepts based on its properties. The aims of classification is to be able to

use the derived model in the prediction of unknown objects. More spacifically,

classification model can be used to classify the unknown class or future object and

develop the classes of the objects in the database in order to get more understanding.

Association is the discovery which use to find the familiarity of

identified patterns that are frequently occur together. Association rule reveals the

associative relationship among the objects [34], for example: Tesco Lotus Phuket

generates an association rule that shows that 50% of time milk is sold with bread and

only 20% of times biscuits are sold with bread.

Clustering is the process to group of similar kind of objects. Clustering

analysis refers to forming group of objects that are very similar to each other. On the

other hands, they are highly different from the objects in other clusters.

In the context of image data, image mining is then suggested. Image

mining is one of a form of data mining which dealing with the extraction of implicit

knowledge, image data relationship or other patterns not explicitly stored in a collection

of images [35]. Image mining consists of multiple components [35] such as: image

analysis, image classification, image indexing, image retrieval and data management.

The component of image mining is presented in Figure 2.16.

30

Figure 2.16 Traditional image mining process

2.4.2 Image Classification

Image classification is a branch of image analysis. It refers to classify an

image by separate or group the similarity properties from the image into the same

category.

With respect to digital image the image classification technique is

applied to assign digital image to classes or categories using the image properties.

Image classification is also widely used in medical image and remote sensing. Pattern

recognitions is another cornerstone in computer science to help image classification

classify an image, p¬attern recognition can be finding the similarities or patterns among

small, decomposed problems that help to classify the complex image. In general, pattern

recognition consists of three steps: (i) spectral pattern recognition (use spatial context

to distinguish between different classes), (ii) spatial pattern recognition (used spectral

reflectance in the different waveband to distinguish between different classes) and (iii)

31

temporal pattern recognition (use variations in image over time to distinguish between

different classes).

Image classification process, the initial step is a collection of image

(known-classes) is used as input data to generate the classification model which can be

used to predict the class of unseen data in the future.

Figure 2.17 Image classification process

Image classification process is presented in Figure 2.17. From the Figure

2.17, image classification is a two sub-process: (i) training and (ii) classification.

During the training process a collection of image data is applied using

pre-processing process to enhance the quality of image and segment the area of interest.

The feature extraction is then applied resulting in a feature vector format. The next step

is a classifier generation, which classification learning methods are applied to construct

the desired classifier. The training data set used in training process, the pre-label record

that typically represented by using an n-dimensional feature vector representation

which in turn a set of attribute values and a class label {i

*1*

*, i*

*2*

*, ..., i*

*n-1*

*, c*

*n*

} where i

i

is an

attribute value and c

*n*

is a class label that c

*n*

∈ C.

Once the classifier has been generated, the second process is

classification. In this process the generated classifier from training process is applied to

the unknown/unseen image data in order to classify/predict the data class.

32

As noted above image classification technique consist two main

processes where in the classification application process need to predict on the usage

of discrete class label.

2.4.3 Classification Learning Methods

As mentioned before the work presented in this thesis is concerned to

image classification. There are nine classification learning methods are presented in this

sub-section: (i) Decision Tree, (ii) Binary Split Tree, (iii) Average One-Dependence

Estimators, (iv) Bayesian Network, (v) Naïve Bayes, (vi) Support Vector Machine, (vii)

Logistic regression, (vii) Sequential Minimal optimization, and (ix) Back Propagation

Neural Network.

**Decision Tree**

Decision Tree [37] is considered as the most popular and widely use for

classification learning methods. A decision tree is a mechanism to classify which is a

recursive partition a collection of instances. Root tree is considered as the top node of

the decision tree. In other words, decision tree is a classifier which instance by sorting

them downs the tree from the root to some leaf node (also known as terminal or decision

nodes). Root node is a direct tree with a node that has no coming edges. Every other

node has exactly one coming edges. An internal or test node is the node that outgoing

edge. In general, test node splits the instance space into two or more sub-spaces

according to a certain discrete function of the input attributes values. Figure 2.18

illustrates the decision tree for the direct mailing response.

Figure 2.18 Response to Direct Mailing Decision Tree.

33

Binary Split Tree

Binary split is the improvement of decision tree. Each node of binary

split tree contains only two values (binary 0-1). The purpose of designing binary split

tree is to store the statistic datasets. The split tree is a subset of decision tree while each

node of a decision tree contain multiple values. The further search in the tree is used by

the split value in the split while the key value is not matched to the search value. The

example of binary split tree is presented in Figure 2.19:

Figure 2.19 Binary Split Tree

**Average One-Dependence Estimators**

Average One-Dependence Estimators (AODE) is presented in this

subsection. AODE is a probabilistic classification learning technique, AODE has

improved from the Naïve Bayesian classifier [38] by addressing the problem of

attribute-independence. For example, in the class of y, which has a set of features x

1

,...,

x

n

, AODE can be applied to find the probability of each class y by using the following

equation:

P̂(y|x

1

,...,x

n

) =

∑ ∑ i:1≤i≤n ∑ ∧ F(x i

)≥m

P̂(y,x

i

) ∏ n

j=1 P̂(x

i

|y,x

i

)

y

′

∈Y

i:1≤i≤n ∧ F(x

i

)≥m

P̂(y′,x

i

) ∏ n

j=1 P̂(x

i

|y′,x

i

)

2.3

34

where

P̂ is an estimate of P

F is the frequency

m is a user specified minimum frequency. Bayesian Network

Bayesian Network (BN) or Probabilistic Networks (PNs) is discussed in

this subsection. BN is consider as a graphical probability model which can be used for

reasoning and the decision making in uncertainty [39]. In addition, the BN have been

applied as a directed acyclic graph (DAG) and each node of BN represents a domain

variable or dataset attribute. BN formally work depend on the Bayes rule while the The

Bayes’ rule can be written as the equation below:

Assume, there are Ai attributes where i= 1,...,n, and take value a

i

where

i= 1,...,n . Assume there is C as class label attribute and U= (a

1

,..., a

n

) as unclassified

test instance. Hence, U will be classified into class C based on Bayes rule is represented

as:

P(C|U) = argmaxP(U|C) 2.4

**Naïve Bayes**

Naïve Bayes has been widely implemented in clustering and

classification. Naïve Bayes sometime called idiot Bayes, simple Bayes or Bayes

classifier. The Bayes theorem is used in Naïve Bayes for prediction both in

classification and clustering. The Bayes theorem is shown as the equation (6) below:

P =

p(d| P(d)

c

i

) p(c

i

)

2.5

where

p(c

i

|d ) is the probability of instance d being in class c

i,

p(d| c

i

) is probability of generating instance d given class c

i,

p( c

i

) is the probability of occurrence of class c

i

p(d) = probability of instance d occurring.

35

**Support Vector Machine**

Support Vector Machine (SVM) is a popular classification method that

has been widely applied for the classification task. The main used of SVM is to separate

instances of two classes by constructing an N-dimensional hyperplane amount of two

training sample classes in the feature set for the most optimal way[40]. SVM classifiers

formally divided into two groups: (i) linear and (ii) non-linear.

• Linear Classification

In linear classification, the SVM can be divided into two

types of classification: (i) linearly separable case, and (ii)

non-linearly separable case. In the linearly separable case,

SVM with the training data x

i

, y i

, y i

∈-1,+1, i = 1,...,n, can be

written as the equation below:

{

x x i

i

.ω + b ≥ +1 for y

i

= +1

2.6

For non-linearly separable, the SVM is defined as below

equation:

{

.ω + b ≤ −1 for y

i

= −1

x

i

.ω + b ≥ +1 − ξ

i

for y

i

= +1 x

i

.ω + b ≤ −1 + ξ

i

for y

i

= −1 ξ

i

2.7 ≥ 0, i = 1,n

• Non-Linear Classification

In non-linear SVM classification, SVM is defined as:

f(x) = ∑ n

i=1 8

α

i

y

i

P(x

i

,x) + b

2.8

where

*n*

*s*

is the number of support vector.

α is non-negative Lagrange multipliers

P (x, y) is Polynominal of degree m: k(x,y)=(x.y+1)m

**Logistic Regression**

Logistic regression is a statistical method which used to analyse a

dataset. The dataset can be one or more independent variables which determine an

36

outcome. The outcome is measured with a dichotomous variable (dependent variable

normally is binary or dichotomous). The purpose of logistic regression is to discover

the best fitting model for evaluating the relationship between a set of independent

variables (predictor) and the dichotomous characteristic of interest (Outcome variable).

Logistic regression provides the formula to predict a logit transformation of the

probability as the Equation 2.9, where p is the probability of presence of the

characteristic of interest.

*logit(p) = b*

*0*

*+ b*

*1*

*X*

*1*

*+ b*

*2*

*X*

*2*

*+ b*

*3*

*X*

*3*

*...+... b*

*k*

*X*

*k*

2.9

**Sequential Minimum Optimization**

Sequential Minimal optimization (SMO) is a computer algorithm which

is used to solve the quadratic programming (QP) problem that happen during the

training of support vector machines (SVM). In order to solve the SVM QP problem,

SMO decomposes SVM QP problem into QP sub-problems then solves the smallest

possible optimization problem which involves two Lagrange multipliers, at each step.

In other words, SMO algorithm can search with feasible region of two kind or type of

problem then maximize the objective function. In other words, by using Lagrangian,

this QP problem can be converted into a dual where the objective function Ψ is solely

dependent on a set of Lagrange multiplier α

i

. The equation of Lagrange is presented in

Equation 2.10:

minΨ (α⃗) ≡ ∑ N

i=1

α

i

− 1 2

∑ i,j

α

i

α

j

y

i

y

j

x

i

x

j 2.10

where

N is the number of training examples

**Back Propagation Neural Network**

Back Propagation Neural Network or Back Propagation is a computer

algorithm/system which sets the model design as the human brain and nervous system.

Back propagation is a mechanism used in artificial neural networks to calculate a

gradient that is needed in the calculation of the weights to be used in the network. Back

37

propagation is commonly used to train deep neural networks,[41] a term referring to

neural networks with more than one hidden layer. Figure 2.20 is illustrated the general

mechanism of Neural Network that has Back propagation to train in deep neural

network:

Figure 2.20 The example of Neural Network mechanism

In the Neural Network (NN) consists of three layers: (i) input, (ii)

hidden, and (ii) output. For Back propagation is use to train in neural for learning. In

other word, in the case of learning in NN, back propagation is commonly applied by

the gradient descent optimization algorithm in order to adjust the weight of neurons by

measuring the gradient of the loss function. This method is sometime known as

backward propagation of errors, because the error is measured at the output layer and

distributed back through the network layers.

**2.5 Feature Selection Methods**

In this section feature selection is described. Feature selection consider

as the main process in the research study in order to make a feature vector which is very

important for the classification process. The main use of feature in the research study

is to reduce data dimensionality for making the feature vector. In the study, there are

five feature selection techniques are applied, including: (i) Correlation-based Feature

Selection (CFS), (ii) Chi-Squared, (iii) information gain, (iv) Gain Ration, and (v)

Relief feature selection.

**1. Correlation-based Feature Selection**

38

Correlation-based feature selection is a techniques used to reduce the

feature space that applied a heuristic search for evaluation the worth of feature subsets.

In work [56] and [57] reported that CFS used the heuristic search for calculation the

evaluation of feature subsets that based on the hypothesis “Good feature subsets contain

features highly correlated with the classification, yet uncorrelated to each other”. In

order to reduce the feature space, CFS used Symmetric Uncertainty. The fooling

equation is defined as the Symmetric Uncertainty equation of two nominal attributes A

and B:

U(A,B) = 2

H(A)+H(B)−H(A,B) H(A)+H(B)

2.10

where

H is the function of entropy

With the reference to Equation 2.10, CFS can be forms as the equation below:

CFS =

∑ n

i=1

U(A

i

,C)

2.11 √ ∑ ∑ U(A

i

where

C is the class of the feature

(A

i

n i=1

n j=1

,A

j

)

,A

j

) is the indicates a pair of attributes

**2. Chi-squared**

Chi-squared is used to measure the relationship of dependency between

a feature and a class [58]. Chi-squared is represented as χ2 symbol and defined as the

equation below:

χ2 = ∑ c

i=1

∑

r j=1

( O

ij

− E

ij

)

2

E

ij

2.12

where

*O*

*ij*

represent as the observed frequency

*E*

*ij*

represent as the expected frequency

39

**3. Information Gain**

Information gain or IG is considered as one of the popular feature

selection techniques. The implementation of IG is for selecting the test attribute at each

node. IG can measure the number of information in bits to the class prediction while

the available information present as a feature and class distribution [59]. For instant, if

IG of a feature s directly related to a collection aspects B, thus IG can be represented

as the equation below:

IG(B,s)∑

v∈Values(s)

|B |B

v

s

|

Entropy(B

v ) 2.13

where

Values(s) presents as the set of all possible feature values s

|B

*v*

|

| presents as the cardinality to the subset of class related to feature s

*|B t*

| presents as the cardinality to the set of aspects belonging to feature s

**4. Gain Ratio**

Gain ration (GR) is presented in this sub-section, an extension of IG is

considered as gain ration technique. The implementation of decision tree in IG for one

reason is to select the test attribute of each node [60]. Thus, the applying of GR to make

IG better performance by choosing an attribute by token number and size of branches.

In other words, GR is implemented to IG in order to reduce the bias of IG on each

branch. GR is written as the equation below:

GR(attribute) =

Entropy(attribute) Gain(attribute)

2.14

**5. Relief**

Relief feature is considered as the last technique is presented in this sub-

section. Relief first introduce in work [61]. Relief is the weight based algorithm that the

related features are consider as the one who has a better distinction between the classes

[62]. In order to understand the relief, a sample of dataset is selected, the nearest

40

neighboring sample that belongs to the same class is called ‘Near-hit’. On the contrary,

the nearest neighboring sample that belongs to the opposite class is called ‘Near-miss’.

The relief measure these two weights: (I) Near-hit and (ii) Near-miss. In addition, all

reiterations of M times, relief takes the feature vector of X that belong to random sample

of Near-hit and Near-miss. After M iteration has gone, relief separates each item of the

weight vector by M, thus, the relevance vector is created. Finally, the selected feature

gets from the features that have relevance value greater than a threshold T value.

**2.6 Deep Learning using Convolutional Neural Network**

The implementation of deep learning in medical image analysis is

presented in this sub-section. Deep learning is one of the mechanism in machine

learning which is directed to the algorithm of brain structure and function called

artificial neural networks. In other words, deep learning is considered as a technique

developed from the artificial neural network which was inspired from human brain

neurons connected system. The application of deep learning has been widely use in

computer vision technology including: image classification, object detection, and

image segmentation. Deep learning model is known as the well-known applications of

deep learning, of those deep learning model is Convolutional Neural Network (CNN)

[42]. CNN is a deep learning techniques that used for image classification. With the

respect to the purposed of study, CNN have been applied to various type of medical

image. For example, the latest study have been shown that deep learning are very

efficient to medical image approach [8, 43, 44, 45, 46]. The Implementation of CCN to

brain MR image for segmented brain into a number of classes is presented in [47]. In

work [48] presented the brain tumors segmentation with MR image with small 3x3

kernel of CNN. For knee application, in work [49] have presented the knee tibia

cartilage segmentation from MR image scan and tested on 114 unseen scans. The

further information of CNN used with respect to work in this thesis is discussed in

Chapter 6.

41

**2.7 Evaluation and Measurement**

To evaluate the classification performance for the generated classifier,

the evaluation measure is applied. With the respect to the work presented in this thesis,

the classification performance was recorded in term of (i) Accuracy (AC), (ii)

Sensitivity (SN), (iii) Specificity (SP), (iv) Precision (PR), and (v) F-Measure (FM).

The evaluation measurement is discussed further detail as follow. The confusion matrix

and some evaluation measures calculating from confusion matrix is presented in 2.7.1.

The basic idea of AUC is discussed in 2.7.2.

2.7.1 Confusion Matrix

The confusion matrix is used to measure how well of applied data

mining algorithm perform on a given data set. It can help to find out which data mining

algorithm give the best performance or the worst performance. The confusion matrix is

performed about the predict class (predict value) and actual class (actual value), Table

2.2 presents the confusion matrix.

Table 2.2 Confusion Matrix

**Predicted Class**

True False

**Actual**

**Class**

True TP TN

False FP FN

From the Table 2.2, TP stands for True Positive rate used to measure the

proportion of correctly identified positive test records. TN refers to the True Negative

which used to measure the proportion of correctly identified negative test records. The

FP is False Positive rate which used to measure the proportion of incorrectly identified

positive test records and FN stands for False Negative which use to measure the

proportion of incorrectly identified negative test records. There are a number of

42

evaluation measures based on the confusion matrix including: (i) Accuracy, (ii)

Sensitivity, (iii) Specificity, (iv) Precision, and (v) F-measure. The equation for each

measure is presented in the Equation 2.11 to Equation 2.15, respectively.

Accuracy =

TP+TN+FP+FN TP+FN

2.15

Sensitivity =

(TP+FN)∗100 TP

2.16

Specificity =

(TN+FP)∗100 TN

2.17

Precision =

TP+FP TP

2.18

F − measure = 2

Precision+Recall Precision∗Recall

2.19

where Recall =

2.20

2.7.1 Area Under Curve

Area under curve (AUC) is considered as the best method for indicating

the overall quality of the classifier [36]. The AUC can be calculated from Receiver

operating characteristic (ROC) graph. ROC graph is a two-dimensional plot of the TP

rate as the y-axis and the FP rate as the x axis [36]. The ROC graph is shown in Figure

2.21.

TP TP+TN

43

Figure 2.21 The ROC graph with five discrete classifiers [36]

From Figure 2.21 the lower left point (0, 0) is the point that a classifier

perform no false positive errors but also gets no true positives. On the contrary, the

upper right point (1, 1) is the point that a classifier perform full false positive errors but

also gets full true positives. Thus, amounts five classifiers A-E, D’s perform the best

one due to the full of true positive rate.

AUC refers to the area under the ROC curve. Figure 2.22 presents the

AUC in ROC graph which the random value from 0.0 to 1.0 in probabilistic classifier.

Figure 2.22 Area Under ROC Curve (AUC) [36]

44

**2.8 The Comparison of related work**

This section the comparison of previous or related works, typically OA

detection, is discussed in this section. Table 2.3 presents the comparison of related

works in this research study:

Table 2.3 The comparison of related works

**Previous**

**work**

**Image**

**Types**

**Osteoarthritis**

**detection**

**Osteoarthritis**

**Stage Detection**

**Accuracy/result,**

**methodology**

Sanjeevakum

ar Kubakaddi

[18]

MRI Just classified

(KOA,

Normal and

Doubtful

KOA)

Calculate the thickness of

cartilage (Shape analysis).

Mahima

Shanker

Pandey [9]

X-ray Detect by

shape analysis

Calculate the thickness of

cartilage with the rate of

result is higher than 65%

(Shape analysis).

Lior Shamir

[10]

X-ray Bone texture of

cartilage

by comparing the value

of each stage and

another prediction use

the values of predicted

example to lower than

the certain value.

45

Table 2.3 The comparison of related works (Continue)

Shivanand

[11]

The result provides

87.92 % by using

different features

calculations such as

shape, statistical, first-

four moment texture and

Haralick texture.

Chao Jin

[50]

X-ray OA/ Non OA

detection by

using

different

technique of

classification

Infrare

OA/Non OA

The result provides

d

by extraction

85.49% of accuracy rate,

the patella-

85.72% of sensitivity

centering of

and 85.51% of

knee feature.

specificity buy using

SVM classifier.

**2.9 Summary**

In conclusion, many kinds of medical images have been applied to

diagnose OA disease with image processing technique as mention above. As an

example [9-18] have been applied the to various of medical image to analyse OA/non-

OA and OA stage using image classification and image procession technique. To

summarise of this chapter is to tell a brief of each previous work had implemented in

image processing technique such as image enhancement, image segmentation, feature

extraction, feature selection, evaluation and measurement. Some techniques for image

processing. There were some brief study for machine learning algorithms for example,

46

neural network, decision tree and logistic regression could be used for classification

process. The detailed of each process work in medical X-ray image analysis is discussed

in Chapter 3.

47