Application of Machine Learning Algorithms On Benchmark Medical Datasets

Project Report submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology.

in

Electronics and Communication Engineering

Under the Supervision of

Dr. Jitendra Virmani

By

Archit Mittal	111073
Pulkit Saxena	111104
Yugander Kishan Singh	111113

to

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNAGHAT

[Approved by UGC under section 2f of UGC act]

Table of Content

CERTIFICATE

This is to certify that project report entitled **"Application of Machine Learning Algorithms on Benchmark medical datasets"** submitted by **Archit Mittal (111073), Pulkit Saxena (111104), Yugander Kishan Singh (111113)** in partial fulfillment for the award of degree of Bachelor of Technology in Electronics And Communication Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

_____ _____

Date: **Dr. Jitendra Virmani**

Assistant Professor (Senior Grade)

ACKNOWLEDGEMENT

We wish to express our deep sense of gratitude to our guide, **Dr. Jitendra Virmani** for his useful suggestions, guidance and support during the project, helping us in completing our project work with improvisation.

The zeal to accomplish the task of formulating the project could not have been realized without the support and cooperation of our parents and faculty members of ECE Department. We sincerely thank Prof. Dr. T.S. Lamba (Dean A&R) and Prof. Dr. Sunil V. Bhooshan (HOD, ECE) for providing the opportunity to undertake the project.

Words are inadequate in offering our thanks to our panel members for their encouragement, support and suggestions which helped us in improving our work and overcoming our shortfalls

(Archit Mittal) (Pulkit Saxena) (Yugander Kishan Singh)

Date:

ABSTRACT

The diagnosis of diseases in most cases depends on a complex combination of clinical and pathological data; this complexity leads to excessive medical costs affecting the cost of medical care. If we look at statistics from WHO, one third of population is suffering from either diabetes or heart disease. Among all diseases heart related disease is found to be the leading cause of death in both males and females and leading in case of female.Computation techniques are often applied for understanding biological phenomena from medical data. For example the discovery of biomarkers in heart disease is one of the key contributions using computational techniques. This process involves the development of predictive model and integration of different types of data and knowledge for diagnostic purposes.

For developing computational techniques related to diseases like heart or diabetes data mining has played an important role in research. To find the hidden medical information from different expression between the healthy and diseased individual in the existed clinical data is a noticeable and powerful approach in study of disease classification. Statistics and machine learning are two main approaches which have been applied to predict the status of disease based on the expression of clinical data.

The diagnosis of diseases in most cases depends on a complex combination of clinical and pathological data. Because of this complexity, there exists a significant amount of interest among clinical professionals and researchers regarding the efficient and accurate prediction of diseases.

In this project, a disease diagnosis system is develop that can assist medical professionals in predicting disease status based on the clinical data of patients. Our approach includes following steps. Firstly, select important clinical features, then develop different algorithms for extraction of feature(in case of imaging database) and then classifying the diseases on the bases of various clinical features, texture and shape of abnormalities that

can be extracted through various tests and find out accuracy of prediction. Finally, develop a user friendly disease prediction system.

In this project, we have studied the various features related to Diabetes, breast cancer and heart disease and with the help of clinical data related to the features of these diseases we have used algorithms like PNN (Probabilistic Neural Network), kNN (K Nearest Neighbor), NN (Neural Network), and SVM (Support Vector Machine) and noted the accuracies obtained from various algorithms to get idea about the efficiencies of algorithms used.

Develop a new and more efficient computer aided diagnosis classification method in digitized mammograms using Smooth Support Vector Machine (SSVM), which performs benign-malignant classification on statistical data based and image based region of interest (ROI) that contains mass, then combine to form a new diagnosis system . The major mammographic characteristics for diagnosis classification is texture, shape and some statistical data. SSVM exploits this important factor to classify it into benign or malignant. The statistical textural features used in characterizing the masses are mean, standard deviation, entropy, skewness and kurtosis and the shape features used in characterizing the masses are area, parameter, length, width, roundness. The main aim of this method and this system is to increase the effectiveness and efficiency of the classification process in an objective manner to reduce the numbers of false-positive of malignancies.

List of Figures

List of Tables

List of Symbols and acronyms

1) CAD	-	Computer Aided Diagnosis
2) SSVM	-	Smooth Support Vector Machine
3) NN	-	Neural Network
4) KNN	-	K- Nearest Neighbor
5) PNN	-	Probabilistic Neural Network
6) MIAS	-	Mammographic Image Analysis Society
7) TEM	-	Texture Energy Measurement
8) ROI	-	Region of Interest
9) ANN	-	Artificial neural network
10) BI-RADS	-	Breast Imaging-Reporting and Data System
11) GDM	-	Gestational diabetes mellitus
12) DM	-	Diabetes mellitus
13) Diastolic BP	-	Diastolic blood pressure
14) Triceps SFT	-	Triceps skin fold thickness
15) BMI	-	Body mass index
16) DPF	-	Diabetes pedigree function

CHAPTER 1

INTRODUCTION

The diagnosis of diseases in most cases depends on a complex combination of clinical and pathological data; this complexity leads to excessive medical costs affecting the cost of medical care. Computation techniques are often applied for understanding biological phenomena from medical data. This process involves the development of predictive model and integration of different types of data and knowledge for diagnostic purposes.

For developing computational techniques related to diseases like heart or diabetes data mining has played an important role in research. To find the hidden medical information from different expression between the healthy and diseased individual in the existed clinical data is a noticeable and powerful approach in study of disease classification. Statistics and machine learning are two main approaches which have been applied to predict the status of disease based on the expression of clinical data.

1.2 Introduction to Datasets

In this project we have used the following datasets:

1.2.1 Diabetes

Diabetes mellitus (DM), commonly referred to as diabetes, is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can cause many complications. Serious long-term complications include cardiovascular disease, stroke, kidney failure, foot ulcers and damage to the eyes.

Prevention and treatment involves a healthy diet, physical exercise, not using tobacco and being a normal body weight. Blood pressure control and proper foot care are also important for people with the disease. Type 1 diabetes must be managed with insulin injections. Type 2 diabetes may be treated with medications with or without insulin. Gestational diabetes usually resolves after the birth of the baby.

As of 2014, an estimated 387 million people have diabetes worldwide, with type 2 diabetes making up about 90% of the cases. This is equal to 8.3% of the adult population, with equal rates in both

women and men. In the years 2012 to 2014, diabetes is estimated to have resulted in 1.5 to 4.9 million deaths per year. Diabetes at least doubles the risk of death. The number of people with diabetes is expected to rise to 592 million by 203

The classic symptoms of untreated diabetes are weight loss, polyuria (frequent urination), polydipsia (increased thirst), and polyphagia (increased hunger). Symptoms may develop rapidly (weeks or months) in type 1 diabetes, while they usually develop much more slowly and may be subtle or absent in type 2 diabetes.

1.2.1.1 Types of Diabetes

Diabetes mellitus is classified into four broad categories: type 1, type 2, gestural diabetes. The term diabetes, without qualification, usually refers to diabetes mellitus.

Type 1 diabetes mellitus is characterized by loss of the insulin-producing beta cells of the islets of Langerhans in the pancreas, leading to insulin deficiency. This type can be further classified as immune-mediated or idiopathic. The majority of type 1 diabetes is of the immune-mediated nature, in which a T-cell-mediated autoimmune attack leads to the loss of beta cells and thus insulin. It causes approximately 10% of diabetes mellitus cases in North America and Europe. Most affected people are otherwise healthy and of a healthy weight when onset occurs. Sensitivity and responsiveness to insulin are usually normal, especially in the early stages. Type 1 diabetes can affect children or adults, but was traditionally termed "juvenile diabetes" because a majority of these diabetes cases were in children.

Type 1diabetes is partly inherited, with multiple genes, including certain HLA genotypes, known to influence the risk of diabetes. In genetically susceptible people, the onset of diabetes can be triggered by one or more environmental factors, such as a viral infection or diet. There is some evidence that suggests an association between type 1 diabetes and Coxsackie B4 virus. Unlike type 2 diabetes, the onset of type 1 diabetes is unrelated to lifestyle.

Type 2 diabetes mellitus is characterized by insulin resistance, which may be combined with relatively reduced insulin secretion. The defective responsiveness of body tissues to insulin is believed to involve the insulin receptor. However, the specific defects are not known. Diabetes mellitus cases due to a known defect are classified separately. Type 2 diabetes is the most common type.

Type 2 diabetes is due primarily to lifestyle factors and genetics. A number of lifestyle factors are known to be important to the development of type 2 diabetes, including obesity (defined by a body mass index of greater than thirty), lack of physical activity, poor diet, stress, and urbanization.

Dietary factors also influence the risk of developing type 2 diabetes. Consumption of sugar-sweetened drinks in excess is associated with an increased risk. The type of fatsin the diet is also important, with saturated fats and trans fatty acids increasing the risk and polyunsaturated and monounsaturated fat decreasing the risk. Eating lots of white rice appears to also play a role in increasing risk. A lack of exercise is believed to cause 7% of cases.

Table 1.1 Comparison of type 1 and 2 diabetes[12]

Comparison of type 1 and 2 diabetes[12]		
Feature	Type 1 diabetes	Type 2 diabetes
Onset	Sudden	Gradual
Age at onset	Mostly in children	Mostly in adults
Body size	Thin or normal	Often obese
Ketoacidosis	Common	Rare
Autoantibodies	Usually present	Absent
Endogenous insulin	Low or absent	Normal, decreased or increased
Concordance in identical twins	50%	90%
Prevalence	~10%	~90%

1.2.1.2 Gestational diabetes

Gestational diabetes mellitus (GDM) resembles type 2 diabetes in several respects, involving a combination of relatively inadequate insulin secretion and responsiveness. It occurs in about 2–10%

of all pregnancies and may improve or disappear after delivery. However, after pregnancy approximately 5–10% of women with gestational diabetes are found to have diabetes mellitus, most commonly type 2. Gestational diabetes is fully treatable, but requires careful medical supervision throughout the pregnancy. Management may include dietary changes, blood glucose monitoring, and in some cases insulin may be required.

1.2.1.3 Dataset Information

For diabetes, we have taken the Pima Indian Diabetes dataset taken from UCI Machine Learning Respiratory Datasets. There are 768 instances; all are women which are mostly pregnant.

(a) Feature Identification and Categorization:-Attributes are usually described by a set of corresponding values. Features described by both numerical and symbolic values can be either discrete (categorical) or continuous. Discrete features concern a situation in which the total number of values is relatively small (finite), while with continuous features the total number of values is very large (infinite) and covers a specific interval (range). The following attributes can be gathered from the data set :

1) Pregnant: Number of times of pregnant
2) Plasma-Glucose: Plasma glucose concentration measured using a two-hour oral glucose tolerance test. Blood sugar level.
3) Diastolic BP: Diastolic blood pressure (mmHg)
4) Triceps SFT: Triceps skin fold thickness (mm)
5) Serum-Insulin: 2-hour serum insulin (mu U/mt)
6) BMI: Body mass index (w in kg/h in m)
7) DPF: Diabetes pedigree function
8) Age: Age of the patient (years)
9) Class: Diabetes onset within five years (0 or 1)

importing samples of the Pima Indian Data Set, changing default attribute titles, and renaming the values of attribute Class from (0, 1) to (No, Yes), one can obtain a proper categorized attribute table.

(b) Short Description

1) <u>Plasma Glucose Concentration</u>

It tells about the amount of glucose in blood. Glucose is the primary source of energy for the body's cells, and blood lipids (in the form of fats and oils) .Glucose is transported from the intestines

or liver to body cells via the bloodstream. Glucose is made available for cell absorption via the hormone insulin, produced by the body primarily in the pancreas.

2) *Diastolic Blood Pressure*

The minimum arterial pressure during relaxation and dilatation of the ventricles of the heart when the ventricles are filled with blood. In a blood pressure reading, the diastolic pressure is typically the second number recorded. For example, with a blood pressure of 120/80 ("120 over 80"), the diastolic pressure is 80. By "80" is meant 80 mm Hg (millimeters of mercury).

3) *Triceps Skin Fold Thickness*

A value used to estimate body fat. The triceps skin fold is the width of a fold of skin taken over the triceps muscle. It is measured using skin fold calipers. Use for the measurement of overall body fat weight and density.

4) *2-Hour Serum Insulin*

Insulin Resistance (IR) is a physiological condition in which cells fail to respond to the normal actions of the hormone insulin. The body produces insulin, but the cells in the body become resistant to insulin and are unable to use it as effectively, leading to hyperglycemia. Beta cells in the pancreas subsequently increase their production of insulin, further contributing to hyper insulinemia. This often remains undetected and can contribute to a diagnosis of Type 2 Diabetes or latent autoimmune diabetes of adults.

5) *Pedigree Chart*

A pedigree chart is a diagram that shows the occurrence and appearance or phenotypes of a particular gene or organism and its ancestors from one generation to the next, most commonly humans, show dogs, and race horses.In this the typical lines and split lines (each split leading to different offspring of the one parent line) resemble the thin leg and foot of a crane.

1.2.2 Mammographic Masses

Breast cancer is cancer that develops from breast tissue. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, or a red scaly patch of skin. In those with distant spread of the disease, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin.

Risk factors for developing breast cancer include obesity, lack of physical exercise, drinking alcohol, hormone replacement therapy during menopause, ionizing radiation.

Mammography is the process of using low-energy X-rays (usually around 30 kVp) to examine the human breast and is used as a diagnostic and a screening tool. The goal of mammography is the early detection of breast cancer, typically through detection of characteristic masses and/or micro calcifications.

Like all X-rays, mammograms use doses of ionizing radiation to create images. Radiologists then analyze the images for any abnormal findings. It is normal to use lower-energy X-rays (typically Mo-K) than those used for radiography of bones

A mammogram may show something suspicious, but by itself it can't prove that an abnormal area is cancer. If a mammogram raises a suspicion of cancer, a tissue sample from the suspicious area must be removed and looked at under the microscope to find out if it is cancer.

1.2.2.1 Dataset Information

For mammographic masses, we have taken the mammographic masses dataset taken from UCI Machine Learning Respiratory Datasets. There are 961 instances.

(a) Feature Identification and Categorization

There are 6 attributes in total (1 goal field, 1 non-predictive, 4 predictive attributes).

1) *BI-RADS assessment*: 1 to 5 (ordinal, non-predictive!)
2) *Age:* patient's age in years (integer)
3) *Shape:* mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
4) *Margin:* mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 speculated=5 .
5) *Density:* mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
6) *Severity:* benign=0 or malignant=1 (binominal, goal field!)

Missing Attribute Values:

- Age: 5
- Shape: 31
- Margin: 48
- Density: 76
- Severity: 0

(b) Short Description

1) *BI-RADS:* It is an acronym for Breast Imaging-Reporting and Data System, a quality assurance tool originally designed for use with mammography. It is further divided into several category:-

   Category 0: Abnormalities may not clearly seen and more tests are needed.
   Category 1:negative.there is no significant abnormality to report.
   Category 2: Benign (non-cancerous) finding.
   Category 3: Suspicious abnormality, Biopsy should be considered.
   Category 4:Highly suggestive of malignancy, appropriate action should be taken.
   Category 5: Known biopsy –proven malignancy –Appropriate action should be taken.
2) *Shape:* If the shape of cancerous tissue is round, oval or lobular the it is benign but if it is Irregular then it is malignant.
3) *Margin:* The margin of breast mass is suspicious for breast cancer when it is ill-defined or speculated.
4) *Density:* If mass density of tissue is high then there is more probability of brest cancer.
5) *Severity:* It's either malignant or benign; It is used as a class identifier.

*A **Benign Tumor:*** is a mass of cells (tumor) that lacks the ability to invade neighboring tissue. These characteristics are required for a tumor to be defined as cancerous and therefore benign tumors are non-cancerous. Also, benign tumors generally have a slower growth rate than malignant tumors

***Malignancy**:* is most familiar as a characterization of cancer. A malignant tumor contrasts with a non-cancerous benign tumor in that a malignancy is not self-limited in its growth, is capable of invading into adjacent tissues, and may be capable of spreading to distant tissues. A benign tumor has none of those properties.

1.2.3 Statlog (Heart)

Heart and blood vessel disease — also called heart disease — includes numerous problems, many of which are related to a process called a atherosclerosis. Atherosclerosis is a condition that develops when a substance called plaque builds up in the walls of the arteries. This buildup narrows

the arteries, making it harder for blood to flow through. If a blood clot forms, it can stop the blood flow. This can cause a heart attack or stroke.

A heart attack occurs when the blood flow to a part of the heart is blocked by a blood clot. If this clot cuts off the blood flow completely, the part of the heart muscle supplied by that artery begins to die. Most people survive their first heart attack and return to their normal lives to enjoy many more years of productive activity. But having a heart attack does mean you have to make some changes. The doctor will advise you of medications and lifestyle changes- according to how badly the heart was damaged and what degree of heart disease caused the heart attack.

Cardiovascular disease is the leading cause of deaths worldwide, though, since the 1970s, cardiovascular mortality rates have declined in many high-income countries. At the same time, cardiovascular deaths and disease have increased at a fast rate in low- and middle-income countries. Although cardiovascular disease usually affects older adults, the antecedents of cardiovascular disease, notably atherosclerosis, begin in early life, making primary prevention efforts necessary from childhood. There is therefore increased emphasis on preventing atherosclerosis by modifying risk factors, for example by health eating, exercise, and avoidance of smoking tobacco.

1.2.3.1 Dataset Information

For mammographic masses, we have taken the heart dataset from UCI Machine Learning Respiratory Datasets. There are 270 instances.

Feature Identification and Categorization

13 attributes in total

1) Age
2) Sex
3) Chest pain types (sharp, dull, burning, aching)
4) Resting Blood Pressure
5) Serum Cholesterol
6) Fasting Blood sugar
7) Resting electrocardiographic results
8) Maximum heart rate achieved
9) Exercise induced angina
10) ST depression induced by exercise relative to rest

11) The slope of the peak exercise ST segment

12) Fluoroscopy

13) Class (1-presence, 2-absence of heart disease)

(a) Short Description

1)*Chest Pain type:-*Chest pain can be of types sharp, dull, burning, aching depending upon cause of its origin.

2)*Resting Blood Pressure:-*In this blood pressure is recorded while patient is resting.

3)*Serum Cholesterol:* -Total amount of cholesterol found in your blood. It is a soft, waxy substance found in the blood and cells of the body.

4) *Fasting Blood Sugar:-*It is a test used for checking heart disease.

5)*Resting Electrocardiographic Results:-*Gives information about the electrical activity of heart.

6) *Exercise induced angina:-*Angina refers to severe pain in chest. It is a test for checking heart disease.

7) *ST Depression:-*ST segment is present in ECG waveform which represents the resting state of heart.

8) *Fluoroscopy:-*It is an imaging technique used for obtaining real time moving images of heart.

1.3 Classifiers

There are different classifiers we use on each and every dataset to that get the best possible result. The different classification techniques which use on different datasets are:-

1.3.1 k-Nearest Neighbor (k-NN):

A nearest-neighbor classification object, where both distance metric ("nearest") and number of neighbors can be altered. The object classifies new observations using the predict method. The object contains the data used for training, so can compute reconstitution predictions.

Just as with the probabilistic methods that we looked at in the previous two lectures, we need a datas et of examples.Each example describes an instance and gives the class to which it belongs. As befor e, we'll assume instances aredescribed by a set of attributevalue pairs, and there is a finite set of class labels **L**. So the dataset comprises examplesof the form **{**A1 = a1, A2 = a2, . . . , An = an**}**, *class* = cl . For a particular instance x, we will refer to x's valuefor attribute A**i** as x.A**i**.In the probabilis tic methods that we looked at, the learning step involved computing probabilities from the dataset.O

nce this was done, in principle the dataset could be thrown away; classification was done using just the probabilities.

In knearestneighbours, the learning step is trivial: we simply store the dataset in the system's memory. In the classification step, we are given an instance q (the quer)whose attributes we will refer to as q.A**i** and we wishto know its class. In kNN, the class of q is found as follows:

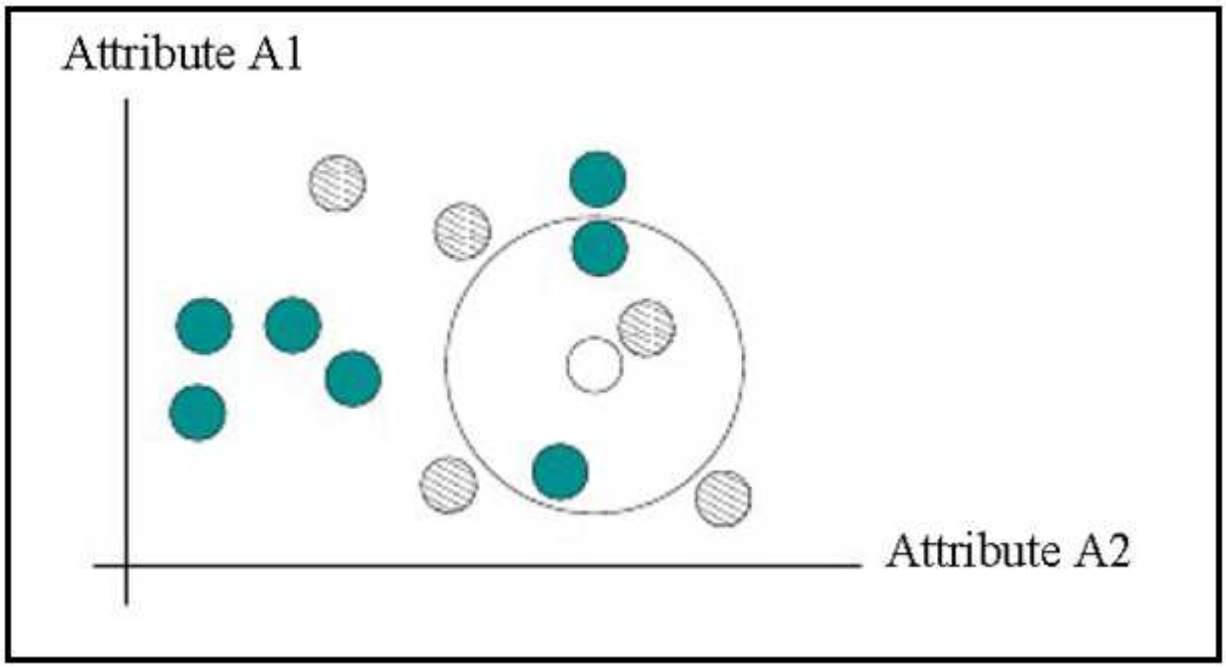

Figure:1.1 kNN

All that remains to do is discuss how distance is measured, and how the voting works.

1.3.2 Probabilistic Neural Network (PNN) :

A probabilistic neural network (PNN) is a feed forward neural network, which was derived from the Bayesian network and a statistical algorithm called Kernel Fisher discriminant analysis. In a PNN, the operations are organized into a multilayered feed forward network with four layers:

- Input Layer
- Hidden Layer
- Pattern Layer/Summation Layer
- Output Layer

(a) *Input layer:*

Each neuron in the input layer represents a predictor variable. In categorical variables, *N-1* neurons are used when there are *N* numbers of categories. It standardizes the range of the values by subtracting the median and dividing by the interquartile range. Then the input neurons feed the values to each of the neurons in the hidden layer.

(b) *Pattern layer:*

This layer contains one neuron for each case in the training data set. It stores the values of the predictor variables for the case along with the target value. A hidden neuron computes the Euclidean distance of the test case from the neuron's center point and then applies the RBF kernel function using the sigma values.

(c) *Summation layer:*

For PNN networks there is one pattern neuron for each category of the target variable. The actual target category of each training case is stored with each hidden neuron; the weighted value coming out of a hidden neuron is fed only to the pattern neuron that corresponds to the hidden neuron's category. The pattern neurons add the values for the class they represent.

(d) *Output layer:*

The output layer compares the weighted votes for each target category accumulated in the pattern layer and uses the largest vote to predict the target category.

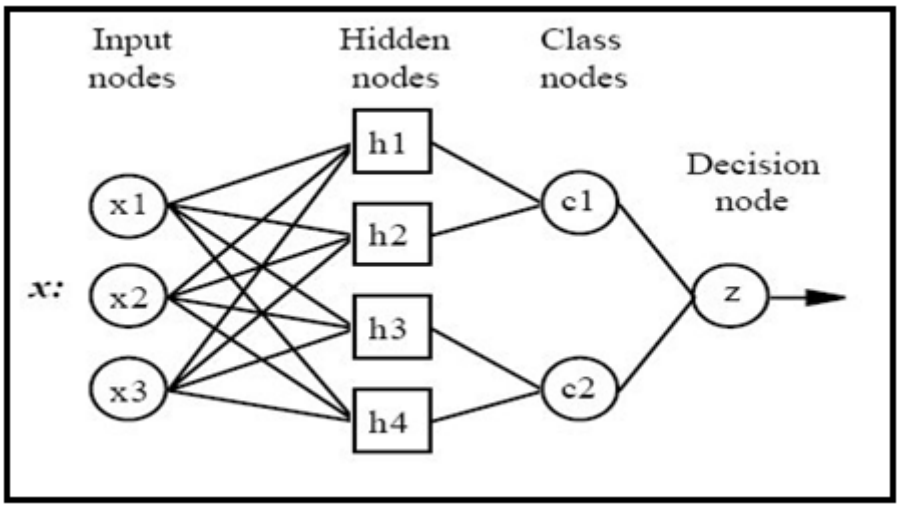

Figure1.2: Layers of Probabilistic neural network

The PNN works by creating a set of multivariate probability densities that are derived from the training vectors presented to the network. The input instance with unknown category is propagated to the pattern layer. Once each node in the pattern layer receives the input, the output of the node will be computed

$$\pi_i^c = \frac{1}{(2\pi)^{n/2}\sigma^n} exp^{-\frac{\left(x-x_{ij}\right)^T (x-x_{ij})}{2\sigma^2}} \quad (1.1)$$

where *d* is the number of features of the input instance *x*, σ is the smoothing parameter, and x_{ij} is a training instance corresponding to category *c*. The summation layer neurons compute the maximum likelihood of pattern *x* being classified into *c* by summarizing and averaging the output of all neurons that belong to the same class

$$p_i(x) = \frac{1}{(2\pi)^{n/2}\sigma^n}\frac{1}{N_i}\Sigma_{i=1}^{N_i} exp^{-\frac{\left(x-x_{ij}\right)^T (x-x_{ij})}{2\sigma^2}} \quad (1.2)$$

where *Ni d*enotes the total number of samples in class *c*.

1.3.3 Support Vector Machine (SVM)

A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are notlinearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function *k(x,y)* selected to suit the problem.The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant. The vectors defining the hyperplanes can be chosen to be linear combinations with parameters α_i of images of feature vectors that occur in the data base. With this choice of a hyperplane, the points x in the feature space that are mapped into the hyperplane are defined by the relation:

$$\alpha_i k(x_i, x) = constant \quad (1.3)$$

Note that if *k(x,y)* becomes small as *y* grows further away from *x*, each term in the sum measures the degree of closeness of the test point *x* to the corresponding data base point x_i. In this way, the sum of kernels above can be used to measure the relative nearness of each test point to the data points originating in one or the other of the sets to be discriminated. Note the fact that the set of points *x* mapped into any hyperplane can be quite convoluted as a result, allowing much more complex discrimination between sets which are not convex at all in the original space.

Given some training data *D*, a set of *n* points of the form

$$D = \{(x_i, y_i) | x_i \in R^p, y_i \in \{-1,1\}\}_{i=1}^{n} \quad (1.4)$$

where the y_i is either 1 or −1, indicating the class to which the point x_i belongs. Each x_i is a p--dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$. Any hyperplane can be written as the set of points x satisfying

$$\omega . x - b = 0 \quad (1.5)$$

where . denotes the dot product and ω the (not necessarily normalized) normal vector to the hyperplane. The parameter $\frac{b}{\|\omega\|}$determines the offset of the hyperplane from the origin along the normal vector ω.

If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin". These hyperplanes can be described by the equations

$$\omega . x - b = 1 \quad (1.6)$$

and

$$\omega . x - b = -1 \quad (1.7)$$

By using geometry, we find the distance between these two hyperplanes is $\frac{2}{\|\omega\|}$, so we want to minimize $\|\omega\|$. As we also have to prevent data points from falling into the margin, we add the following constraint: for each i either

$$\omega . x - b \geq -1 \text{ for } x_i \text{ of the first class} \quad (1.8)$$

or

$\omega.x - b \leq -1$ for x_i of the second. (1.9)

This can be rewritten as:

$y_i(\omega.x - b) \leq -1$ for all $1 \leq i \leq n$ (1.10)

We can put this together to get the optimization problem:

Minimize (in ω,b)

$\|\omega\|$

subject to (for any *i=1,2…,n*)

$y_i(\omega.x - b) \geq 1$ (1.11)

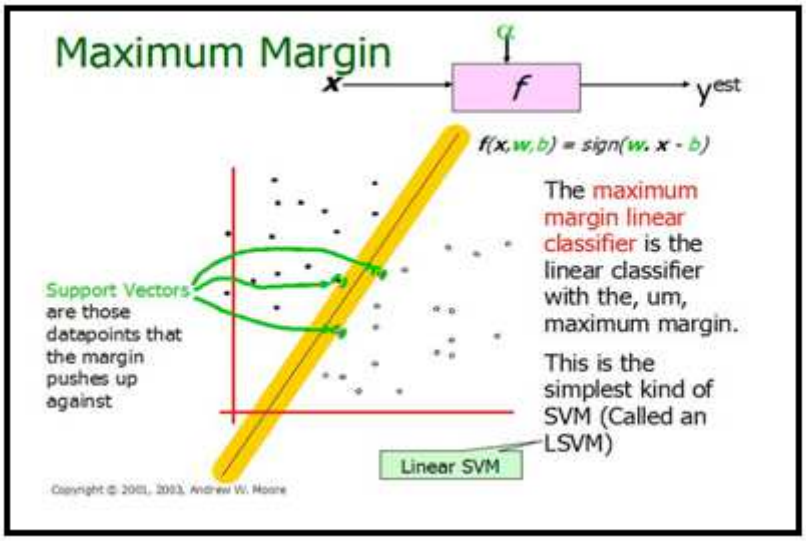

Figure 1.3 Support Vector Machine

1.2.4 Smooth Support Vector Machine:

Smoothing methods, extensively used for solving important mathematical programming problems and applications are applied here to generate and solve an unconstrained smooth reformulation of the support vector machine for pattern classification using a completely arbitrary kernel. We term such reformulation a smooth support vector machine (SSVM). A fast Newton-Armijo algorithm for solving the SSVM converges globally and quadratically. Numerical results and comparisons are given to demonstrate the effectiveness and speed of the algorithm. SSVM is an advanced form of

SVM. Also SSVM gives greater accuracy as compared to other classification techniques that we have used namely NN, PNN, and KNN.

Now we consider the problem of classifying *m* points in the *n*-dimensional real space R^n, represented by the *m*× *n* matrix A, according to membership of each point A_i in the classes 1 or -1 as specified by a given *m*× *m* diagonal matrix D with ones or minus ones along its diagonal. For this problem the standard support vector machine with a linear kernel AA′ [30] [12] is given by the following for some *v>0:*

$$\min_{(\omega,\gamma,y)\epsilon \mathbf{R}^{n+m+1}} ve^{'}y + \frac{1}{2}\omega^{'}\omega$$

$$\text{s.t. } D(A\omega - e\gamma) + y \geq e \qquad (1.12)$$

$$y \geq 0$$

Here$\omega$is the normal to the bounding planes:

$$x^{'}\omega - \gamma = +1$$

$$x^{'}\omega - \gamma = -1 \qquad (1.13)$$

and γdetermines their location relative to the origin. The first plane above bounds the class 1 points and the second plane bounds the class -1 points when the two classes are strictly linearly separable, that is when the slack variable y = 0. The linear separating surface is the plane

$$x^{'}\omega = \gamma \qquad (1.14)$$

midway between the bounding planes (1.13). See Figure 1.5. If the classes are linearly inseparable then the two planes bound the two classes with a “soft margin” determined by a nonnegative slack variable *y*, that is:

$$x^{'}\omega - \gamma + y_i \geq +1, \text{for} x^{'} = A_i \text{and} D_{ii} = +1, \qquad (1.15)$$

$$x^{'}\omega - \gamma + y_i \leq -1, \text{for} x^{'} = A_i \text{and} D_{ii} = -1$$

The 1-norm of the slack variable y is minimized with weight _ in (1). The quadratic term in (1.12), which is twice the reciprocal of the square of the 2-norm distance $\frac{2}{\|\omega\|_2}$between the two bounding planes of (1.13) in the *n*-dimensional space of $\omega \in R^n$for a fixed γ, maximizes that distance, often called the “margin”. Figure 1 depicts the points represented by A, the bounding planes (1.13) with

margin $\frac{2}{\|\omega\|_2}$, and the separating plane (3) which separates A+, the points represented by rows of A with $D_{ii} = +1$, from A−, the points represented by rows of A with $D_{ii} = -1$.

In our smooth approach, the square of 2-norm of the slack variable y is minimized with weight $\frac{v}{2}$ instead of the 1-norm of y as in (1). In addition the distance between the planes (2) is measured in the $(n + 1)$-dimensional space of$(w, \gamma) \in R^{n+1}$, that is $\frac{2}{\|\omega\|_2}$. Measuring the margin in this (n+1)-dimensional space instead of R^n induces strong convexity and has little or no effect on the problem as was shown in [23]. Thus using twice the reciprocal squared of the margin instead, yields our modified SVM problem as follows:

$$\min_{\omega,\gamma,y} \frac{v}{2} y'y + \frac{1}{2}(w'w + \gamma^2)$$
$$\text{s.t. } D(A\omega - e\gamma) + y \geq e \qquad (1.16)$$
$$y \geq 0$$

At a solution of problem (5), y is given by

$$y = \left(e - D(A\omega - e\gamma)\right)_+$$

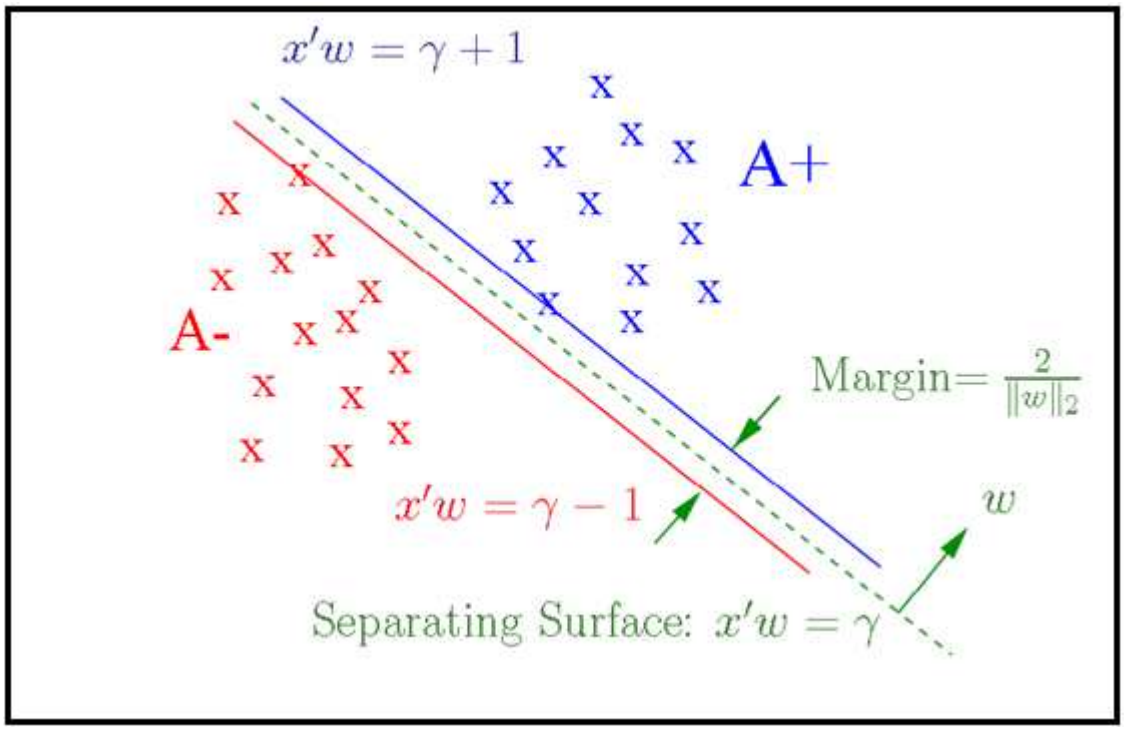

Figure: 1.4 Smooth Support Vector Machine

where, as defined earlier, $(\cdot)_+$ replaces negative components of a vector by zeros. Thus, we can replace y in (5) by $\left(e - D(A\omega - e\gamma)\right)_+$ and convert the SVM problem (5) into an equivalent SVM which is an unconstrained optimization problem as follows:

$$\min_{w,\gamma} \frac{v}{2} \left\| \left(e - D(A\omega - e\gamma)\right)_+ \right\|_2^2 + \frac{1}{2}(\omega'\omega + \gamma^2) \qquad (1.18)$$

This problem is a strongly convex minimization problem without any constraints. It is easy to show that it has a unique solution. However, the objective function in (7) is not twice differentiable which precludes the use of a fast Newton method. We thus apply the smoothing techniques and replace x_+ by a very accurate smooth approximation that is given by $p(x, \alpha)$, the integral of the sigmoid function $\frac{1}{1+\varepsilon^{-\alpha x}}$ of neural networks, that is

$$p(x, \alpha) = x + \frac{1}{\alpha}\log(1 + \varepsilon^{-\alpha x}), \alpha > 0 \qquad (1.19)$$

This p function with a smoothing parameter α is used here to replace the plus function of (1.18) to obtain a smooth support vector machine (SSVM):

$$\min_{(w,\gamma)\in R^{n+1}} \emptyset_\alpha(\omega, \gamma) = \min_{(\omega,\gamma)\in R^{n+1}} \frac{v}{2} \|p(e - D(A\omega - e\gamma), \alpha)\|_2^2 + \frac{1}{2}(w'w + \gamma^2) \quad (1.20)$$

1.2.5 Neural Network (NN):

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well.

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. Artificial neural network (ANN) is a machine learning approach that models human brain and consists of a number of artificial neurons.Neuron in ANNs tend to have fewer connections than biological neurons.Each neuron in ANN receives a number of inputs.An activation function is applied to these inputs which results in

activation level of neuron (output value of the neuron).Knowledge about the learning task is given in the form of examples called training examples

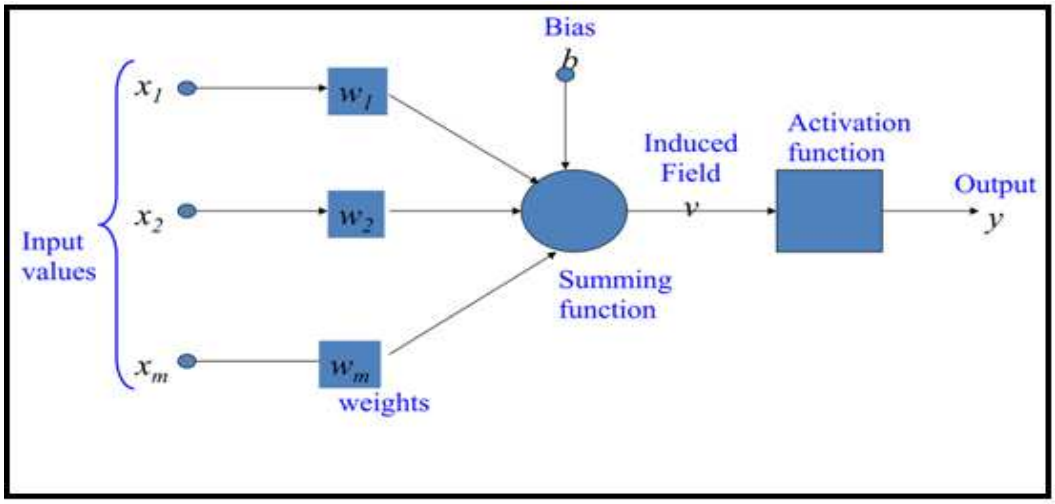

Figure: 1.5 Complete Neural Network with Activation Function.

(a) Bias of a Neuron:-

1) The bias *b* has the effect of applying a transformation to the weighted sum *u*

   $v = u + b$

2) The bias is an external parameter of the neuron. It can be modeled by adding an extra input.

3) *v*is called induced field of the neuron

$$v = \sum_{j=0}^{m} w_j x_j$$
$$w_0 = b \qquad (1.21)$$

4) The choice of activation function φ determines the neuron model.

   Examples:

5) step function: $\varphi(v) = \begin{cases} a \text{ if } v < c \\ b \text{ if } v > c \end{cases}$

6) ramp function: $\varphi(v) = \begin{cases} a \text{ if } v < c \\ b \text{ if } v > d \\ a + ((v-c)(d-c)/(b-a)) \text{ otherwise} \end{cases}$ (1.22)

7) sigmoid function with z,x,y parameters $\varphi(v) = z + \dfrac{1}{1+\exp(-xv+y)}$ (1.23)

8) Gaussian function: $\varphi(v) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right)$

(b) Advantages :

1) *Adaptive learning:* An ability to learn how to do tasks based on the data given for training or initial experience.
2) *Self-Organization:* An ANN can create its own organization or representation of the information it receives during learning time.
3) *Real Time Operation:* ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability.
4) *Fault Tolerance via Redundant Information Coding:* Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage.

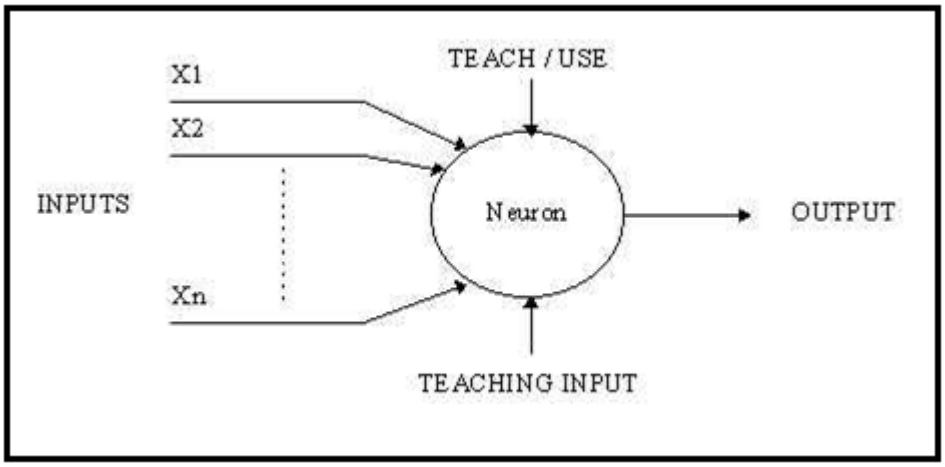

Figure: 1.6 Neural Network

5) After forming the train model by applying the different classifiers on the datasets, then the patient's data is tested with the help of testing data on the network.
6) Then, efficiency of classifier is calculated.

CHAPTER 2

CAD SYSTEM FROM UCI BENCHMARK DATASETS

2.1 Introduction

Computer Aided Diagnosis System for benchmark databases are widely used system**,** in which we distribute the whole dataset into two equal sizes dataset i.e. Training Set and Testing Set. After, distribution of dataset, check for Missing Values. If Missing Values are present in the dataset, apply Missing Value Treatment. Otherwise, the classification methods will not provide accurate results. After Missing Value Treatment, Feature Selection Method should be applied to extract only relevant features, but due to less number of features, Feature Selection Techniques are not advisable. After performing Feature Selection Technique, perform different classification techniques. In this projection, following classification techniques are used:-

1) k- Nearest Neighbor
2) Probabilistic Neural Network
3) Neural Network
4) Support Vector Machine
5) Smooth Support Vector Machine

2.2 Design of CAD Model

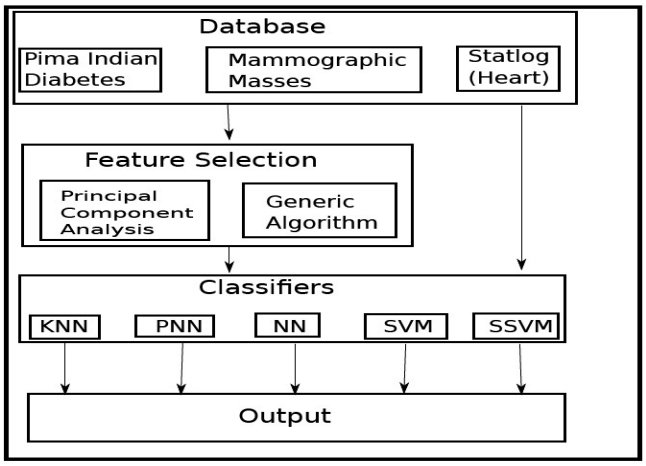

Figure:2.1 CAD Model

2.3 Training And Testing Sets

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, you can minimize the effects of data discrepancies and better understand the characteristics of the model.

After a model has been processed by using the training set, you test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that you want to predict, it is easy to determine whether the model's guesses are correct.

2.4 Steps for Classification

1) Take the dataset from UCI Machine Learning Respiratory Dataset.
2) Randomly distribute the database in excel and divide it equally in two datasets namely Training Dataset and Testing Dataset.

2.5 Missing Value Treatment

We have to treat the dataset with appropriate Missing Value Treatment so that we can eradicate the values which are irrelevant. Missing data are a part of almost all research, and we all have to decide how to deal with it from time to time. There are a number of alternative ways of dealing with missing data

1) *The simplest approach list wise deletion:-*
the most common approach to missing data is to simply omit those cases with missing data and to run our analyses on what remains.
Although list wise deletion often results in a substantial decrease in the sample size available for the analysis, it does have important advantages.
2) *Mean substitution:-*
An old procedure that should certainly be relegated to the past was the idea of substituting a mean for the missing data.
We have really added no new information to the data but we have increased the sample size. The effect of increasing the sample size is to increase the denominator for computing the standard error, thus reducing the standard error**.**

Now, apply different classification technique on training dataset and make the network model, but before we use, we have to study the classification technique.

2.6 Results

The results of different classifiers on different datasets are given below:-

2.6.1 Pima Indian Diabetes:-

TABLE: 2.1 Pima Indian Diabetes Results

Classifier	Confusion Matrix	Accuracy (%)	S_A (%)	S_N (%)
K-Nearest Neighbor	N: 208, 42 A: 49, 85	76.8(K=10)	63.4	83.2
Probabilistic Neural network	N: 161, 39 A: 39, 95	75(Spread=.71)	70.86	64.4
Support Vector Machine	N: 219, 31 A: 64, 70	75.26	77.37	69.31
Neural Network	N: 236, 14 A: 126, 8	63.54	5.97	94.4
Smooth Support Vector Machine	N: 219, 31 A: 42, 92	80.98	87.6	68.6

N: - Normal, A: - Abnormal, S:-Sensitivity.

2.6.2 Mammographic Masses

TABLE:2.2 Mammographic Masses Result

Classifier	Confusion Matrix	Accuracy (%)	S_A (%)	S_N (%)
K-Nearest Neighbour	B: 212, 36; M: 57, 156	79.6(K=3)	73.23	85.48
Probabilistic Neural Network	B: 184, 64; M: 34, 236	78.7(Spread=.50)	74.19	84.03
Support Vector Machine	B: 201, 47; M: 45, 168	80.04	81.70	78.13
Neural Network	B: 181, 67; M: 30, 183	79	85.91	72.9
Smooth Support Vector Machine	B: 206, 42; M: 28, 185	84.82	83.06	68.6

B: - Benign, M: - Malignant, S:-Sensitivity.

2.6.3 Statlog (heart):-

TABLE:2.3 Statlog (Heart) Results

Classifier	Confusion Matrix	Accuracy (%)	S_N (%)	S_A (%)
K-Nearest Neighbour	N: 54, 23 / A: 26, 36	67.8(K=7)	70.13	64.28
Probabilistic Neural network	N: 95, 39 / A: 89, 161	69.2(Spread=.40)	70.89	64.4
Support Vector Machine	N: 60, 14 / A: 14, 45	78.9	81.08	76.27
Neural Network	N: 57, 17 / A: 10, 49	79.7	77.02	83.05
Smooth Support Vector Machine	N: 68, 10 / A: 11, 48	84.67	87.17	81.35

N: - Normal, A: - Abnormal, S:-Sensitivity.

2.6 Outcome

SSVM provides better accuracy than SVM and other classification technique on all the three datasets. So we will perform SSVM classifier for further project.

CHAPTER 3

CAD SYSTEM FROM IMAGING DATABASE (MIAS DATABASE)

3.1 Introduction

Computer-aided detection (CAD) is a recent advance in the field of breast imaging and is designed to improve radiologists' ability to find even the smallest breast cancers at their earliest stages.CAD software uses sophisticated algorithms based on several thousand cases of breast cancer to identify suspicious areas on a mammogram that might warrant close examination. The patient gets a digital mammogram in the usual way. There are no additional steps for patients to undergo. Our radiologist reviews the mammogram without CAD analysis and makes a preliminary interpretation. Once that initial review is done, we apply the CAD software to the mammogram. It highlights areas of potential concern and the radiologist takes another look at them. Because CAD is very sensitive and can detect very subtle abnormalities, it has proven to be particularly useful in mammography exams involving dense breast tissue

With the CAD technology, the radiologist still reviews all aspects of the mammogram and makes the final interpretation. CAD cannot diagnose. It simply serves as a highly valuable double check.

Abnormalities in mammogram can be detected by

1) Texture Feature: - An image texture is a set of metrics calculated in image processing designed to quantify the perceived texture of an image. Image texture gives us information about the spatial arrangement of color or intensities in an image or selected region of an image

2) Shape Based Feature:-The feature which will be retrieve from the shape of the image like area, parameter, length, circularity etc.However, the knowledge will be useless if one can't find it. Face to the substantive and increasing apace images, how to search and to retrieve the images that we are interested in facility is a fatal problem

3.2 Design of CAD Model

The CAD model of the texture feature and shape feature are as follows:-

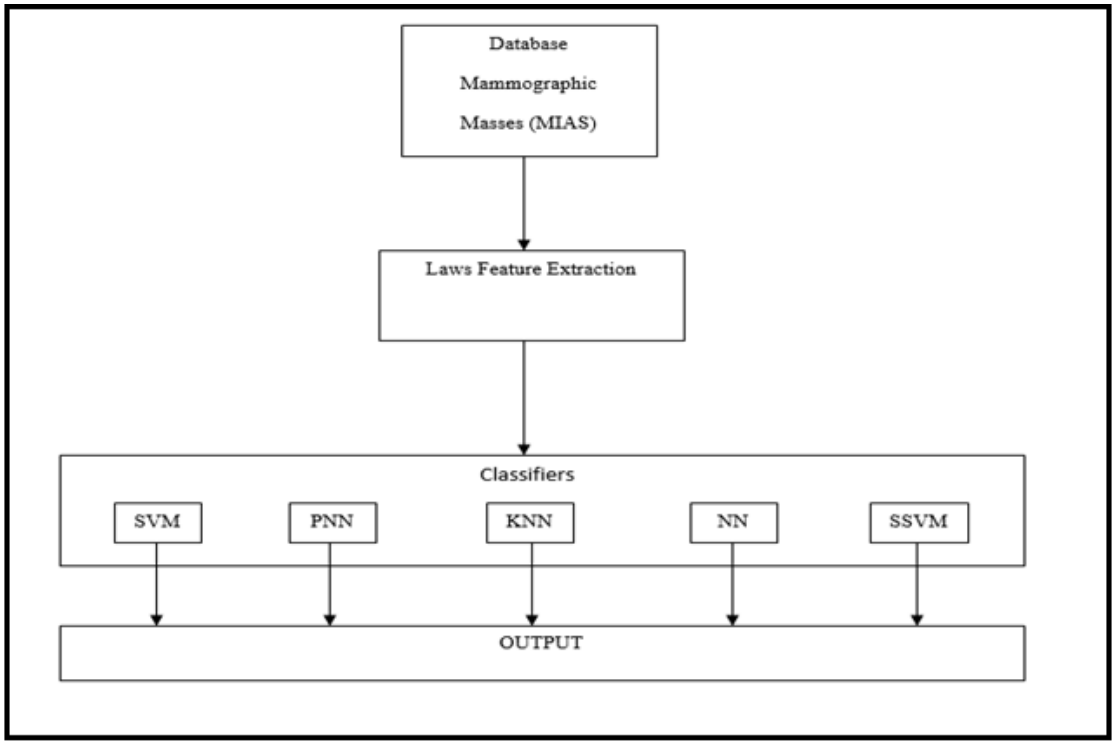

Figure: 3.1 CAD Model 2 (Texture Feature Extraction)

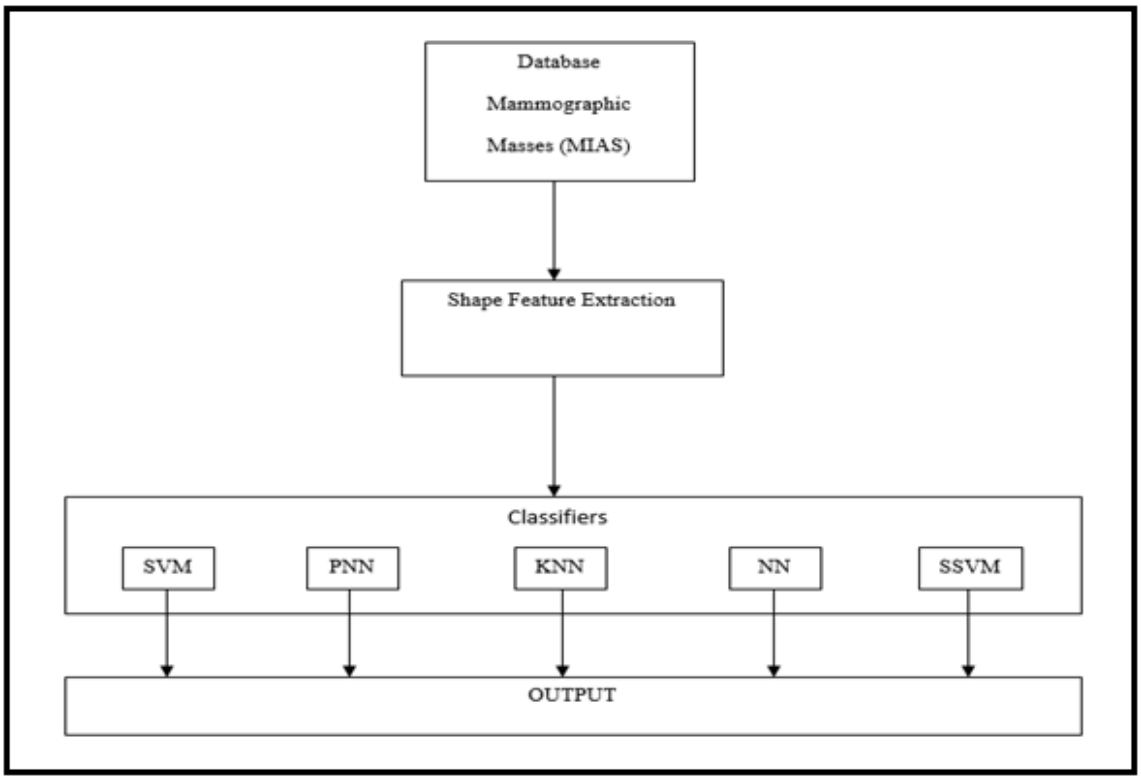

Figure:3.2 CAD Model 3 (Shape Feature Extraction)

3.3 Texture Feature

An image texture is a set of metrics calculated in image processing designed to quantify the perceived texture of an image. Image texture gives us information about the spatial arrangement of color or intensities in an image or selected region of an image.

Image textures can be artificially created or found in natural scenes captured in an image. Image textures are one way that can be used to help in segmentation or classification of images. To analyze an image texture in computer graphics, there are two ways to approach the issue: Structured Approach and Statistical Approach.

(a) Structural Approach: - A structured approach sees an image texture as a set of primitive texels in some regular or repeated pattern. This works well when analyzing artificial textures.

To obtain a structured description a characterization of the spatial relationship of the texels is gathered by using Voronoi tessellation of the texels.

(b) Statistical Approach: - A statistical approach sees an image texture as a quantitative measure of the arrangement of intensities in a region. In general this approach is easier to compute and is more widely used, since natural textures are made of patterns of irregular subelements.Some of the statistical Approaches are:-

1) Edge Detection: - The use of edge detection to determine the number of edge pixels in a specified region helps determine a characteristic of texture complexity. After edges have been found the direction of the edges can also be applied as a characteristic of texture and can be useful in determining patterns in the texture.

*2) Co-occurrence Matrices:-*Co-occurrence matrix captures numerical features of a texture using spatial relations of similar gray tones. Numerical features computed from the co-occurrence matrix can be used to represent, compare, and classify textures.

3) Laws Texture Energy Measures:- Another approach to generate texture features is to use local masks to detect various types of textures.

3.3.1 Laws Texture Energy Measures

Feature extraction is the process of obtaining higher-level information of an image such as color, shape and texture. Texture is a key component of human visual perception. Although there is no strict definition of the image texture, it is easily perceived by humans and is believed to be a rich source of visual information – about the nature and three dimensional shapes of physical objects. Generally speaking, textures are complex visual patterns composed of entities, or sub patterns that have characteristic brightness, color, slope, size, etc. Thus texture can be regarded as a similarity grouping in an image. The local sub pattern properties give rise to the perceived lightness, uniformity, density, roughness, regularity, linearity, frequency, phase, directionality, coarseness, randomness, fineness, smoothness, granulation, etc., of the texture as a whole. For a large collection of examples of textures. There are four major issues in texture analysis:

1) *Feature extraction:* to compute a characteristic of a digital image able to numerically describe its texture properties;
2) *Texture discrimination:* to partition a textured image into regions, each corresponding to a perceptually homogeneous texture (leads to image segmentation);
3) *Texture classification:* to determine to which of a finite number of physically defined classes (such as normal and abnormal tissue) a homogeneous texture region belongs;
4) *Shape from texture:* to reconstruct 3D surface geometry from texture information.

(a) Feature extraction techniques:-

First-order histogram based features

Assume the image is a function f(x,y) of two space variables x and y, x=0,1,…,N-1 and y=0,1,…,M-1. The function f(x,y) can take discrete values i = 0,1,…,G-1, where G is the total number of number of intensity levels in the image. The intensity-level histogram is a function showing (for each intensity level) the number of pixels in the whole image, which have this intensity:

$$h(i) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \delta(f(x,y), i) \qquad (2.1)$$

where $\delta(j,i)$ is the Kronecker delta function

The histogram of intensity levels is obviously a concise and simple summary of the statistical information contained in the image. Calculation of the grey-level histogram involves single pixels.

Thus the histogram contains the first-order statistical information about the image (or its fragment). Dividing the values h(i) by the total number of pixels in the image one obtains the approximate probability density of occurrence of the intensity levels

$$p(i) = h(i)/NM \qquad (2.2)$$

The histogram can be easily computed, given the image. The shape of the histogram provides many clues as to the character of the image. For example, a narrowly distributed histogram indicated the low-contrast image. A bimodal histogram often suggests that the image contained an object with a narrow intensity range against a background of differing intensity. Different useful parameters (image features) can be worked out from the histogram to quantitatively describe the first-order statistical properties of the image. Most often the so-called central moments are derived from it to characterize the texture as defined by Equations below.

Mean: $\mu = \sum_{i=0}^{G-1} ip(i)$ (2.3)

Variance: $\sigma^2 = \sum_{i=0}^{G-1}(i-\mu)^2 p(i)$ (2.4)

Skewness: $\mu_3 = \sigma^{-3}\sum_{i=0}^{G-1}(i-\mu)^3 p(i)$ (2.5)

Kurtosis: $\mu_4 = \sigma^{-4}\sum_{i=0}^{G-1}(i-\mu)^4 p(i) - 3$ (2.6)

Two other parameters are also used, described below.

Energy: $E = \sum_{i=0}^{G-1}[p(i)]^2$ (2.7)

Entropy: $H = -\sum_{i=0}^{G-1} p(i) log_2[p(i)]$ (2.8)

The mean takes the average level of intensity of the image or texture being examined, whereas the variance describes the variation of intensity around the mean. The skewness is zero if the histogram is symmetrical about the mean, and is otherwise either positive or negative depending whether it has been skewed above or below the mean. Thus μ3 is an indication of symmetry. The kurtosis is a measure of flatness of the histogram; the component '3' inserted in (4) normalises μ4 to zero for a Gaussian-shaped histogram. The entropy is a measure of histogram uniformity. Other possible

features derived from the histogram are the minimum, the maximum, the range and the median value.

In the case of visual images, the mean and variance do not actually carry the information about the texture. They rather represent the image acquisition process, such as the average lighting conditions or the gain of a video amplifier. Using images normalised against both the mean and variance can give better texture discrimination accuracy than using the actual mean and the actual variance as texture parameters. Thus images are often normalised to have the same mean, e.g.$\mu = 0$, and the same standard deviation, e.g. $\sigma = 1$.

The Laws method uses filter masks to extract secondary features from natural micro-structure characteristics of the image (level, edge, spot and ripple) which can then be used for segmentation or classification. Laws developed five labeled vectors which could be combined to form two dimensional convolution kernels. When convolved with a textured image these masks extract individual structural components of the image.

The images can be filtered with some specific masks to assess texture properties. These masks are created by combination of different one dimensional kernel vectors.5 types of masks are used: Level(L),Edge(E),Spot(S),Ripple(R),Wave(W).These laws vectors are used in different dimensions like:3,5,7 and 9.

Level - Gaussian - Gives a central weighted local average

Edge -Gradient - responds to row and column step edges

Spot - Log- detects spots.

wave detection.

Ripple - Gabor- detects ripple

Vectors of resolution 3 are-

L3= [1, 2, 1]

E3= [-1, 0, 1]

S3= [-1, 2, 1]

The 2-D masks formed from these vectors are

L3L3	E3L3	S3L3
L3E3	E3E3	S3E3
L3S3	E3S3	S3S3

Vectors of resolution 5 are-

L5= [1, 4, 6, 4, 1]

E5= [-1,-2, 0, 1, 2]

S5= [-1, 0, 2, 0,-1]

W5= [-1, 2, 0,-2, 1]

R5= [1,-4, 6,-4, 1]

The 2-D masks formed from these vectors are

L5L5	E5L5	S5L5	W5L5	R5L5
L5E5	E5E5	S5E5	W5E5	R5E5
L5S5	E5S5	S5S5	W5S5	R5S5
L5W5	E5W5	S5W5	W5W5	R5W5
L5R5	E5R5	S5R5	W5R5	R5R5

Vectors of resolution 7 are-

L7= [1, 6, 15, 20, 15, 6, 1]

E7= [-1,-4,-5, 0, 5, 4, 1]

S7= [-1,-2, 1, 4, 1,-2,-1]

The 2-D masks formed from these vectors are

L7L7	E7L7	S7L7
L7E7	E7E7	S7E7
L7S7	E7S7	S7S7

Vectors of resolution 9 are-

L9=[1,8,28,56,70,56,28,8,1]

E9=[1,4,4,-4,-10,-4,4,4,1]

S9=[1,0,-4,6,0,-4,0,1]

W9=[1,-4,4,-4,-10,4,4,-4,1]

R9=[1,-8,28,-56,70,-56,28,-8,1]

The 2-D masks formed from these vectors are

L9L9	E9L9	S9L9	W9L9	R9L9
L9E9	E9E9	S9E9	W9E9	R9E9
L9S9	E9S9	S9S9	W9S9	R9S9
L9W9	E9W9	S9W9	W9W9	R9W9

L9R9 E9R9 S9R9 W9R9 R9R9

After a series of particular convolution with selected Laws' masks, the outputs are passed to texture energy measurement (TEM) filters for the analysis of the texture property of each pixel. These consisted of a moving nonlinear window operation; every pixel of the image is replaced by comparing the pixel with its local neighborhood based on three statistical descriptors (mean, absolute mean and standard deviation). These descriptors are computed as follows:

$$mean = \frac{\sum_W neighbouring\ pixels}{W} \tag{2.9}$$

$$absolute\ mean = \frac{\sum_W abs(neighbouring\ pixels)}{W} \tag{2.10}$$

$$standard\ deviation = \sqrt{\frac{\sum_W (neighbouring\ pixels - mean)^2}{W}} \tag{2.11}$$

where W is the window size. The operation will lead to the creation of three TEM images corresponding to each statistical descriptor. After the windowing operation, all the obtained images are normalized in order to be presented well as images. Min-max normalization method is utilized in this work.Subsequently, for each normalized TEM (NTEM) image we compute three statistics; absolute mean (ABSM), mean square or energy (MS) and entropy as follows:

$$ABSM = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}|l(x,y)| \tag{2.12}$$

$$MS = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}l(x,y)^2 \tag{2.13}$$

$$Entropy = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}l(x,y)(-\ln l(x,y)) \tag{2.14}$$

where I(x, y) is the pixel value, and M and N are image dimensions.

Steps followed in Laws 'mask analysis are:

1) Convolving the image I(i,j) with each 2-D mask forming a texture image (TI).e.g.

$$\mathrm{TI_{E5E5}} = \mathrm{I_{i,j}} \otimes \mathrm{E5E5} \tag{2.15}$$

2) Normalizing the contrast of texture image.

$$\text{Normalize}\,(\text{TI}_{\text{mask}}) = \frac{\text{TI}_{\text{mask}}}{\text{TI}_{\text{L5L5}}} \qquad (2.16)$$

3) The TIs are passed through Texture Energy Measurement (TEM) filters.

$$\text{TEM}_{\text{i,j}} = \sum_{\text{u}=-7}^{7} \sum_{\text{v}=-7}^{7} [\text{Normalize}(\text{TI}_{\text{i+u,j+v}})] \qquad (2.17)$$

4) By combining 25 TEM descriptors we obtain 15 rotationally invariant TEMs denoted as TR.

$$\text{TR}_{\text{E5L5}} = \frac{\text{TEM}_{\text{E5L5}} + \text{TEM}_{\text{L5E5}}}{2} \qquad (2.18)$$

5) From each TR five statistical parameters are obtained namely: Mean, Standard deviation (SD), Skewness, Kurtosis, Entropy

$$\text{Mean} = \frac{\sum_{\text{i}=0}^{\text{M}} \sum_{\text{j}=0}^{\text{N}} [\text{TR}_{\text{i,j}}]}{\text{M} \times \text{N}} \qquad (2.19)$$

$$\text{SD} = \sqrt{\frac{\sum_{\text{i}=0}^{\text{M}} \sum_{\text{j}=0}^{\text{N}} (\text{TR}_{\text{i,j}} - \text{Mean})^2}{\text{M} \times \text{N}}} \qquad (2.20)$$

$$\text{Skewness} = \frac{\sum_{\text{i}=0}^{\text{M}} \sum_{\text{j}=0}^{\text{N}} (\text{TR}_{\text{i,j}} - \text{Mean})^3}{\text{M} \times \text{N} \times \text{SD}^3} \qquad (2.21)$$

$$\text{Kurtosis} = \frac{\sum_{\text{i}=0}^{\text{M}} \sum_{\text{j}=0}^{\text{N}} (\text{TR}_{\text{i,j}} - \text{Mean})^4}{\text{M} \times \text{N} \times \text{SD}^4} - 3 \qquad (2.22)$$

$$\text{Entropy} = \frac{\sum_{\text{i}=0}^{\text{M}} \sum_{\text{j}=0}^{\text{N}} (\text{TR}_{\text{i,j}})^2}{\text{M} \times \text{N}} \qquad (2.23)$$

3.4 Shape Based Feature Detection

Visual information plays an important role in our society, it will play an increasinglypervasive role in our lives, and there will be a growing need to have these sources processed further. The pictures or images are used in many application areas like architectural and engineering design, fashion, journalism, advertising, entertainment, etc. Thus it provides the necessary opportunity for us to use the abundance of images. However, the knowledge will be useless if one can't find it. Face to the substantive and increasing apace images, how to search and to retrieve the images that we are

interested in facility is a fatal problem: it brings a necessity for image retrieval systems. As we know, visual features of the images provide a description of their content. Content-based image retrieval, emerged as a promising mean for retrieving images and browsing large images databases. Content-based image retrieval has been a topic of intensive research in recent years. It is the process of retrieving images from a collection based on automatically extracted features from those images. Efficient shape features must present some essential properties given in the figure:

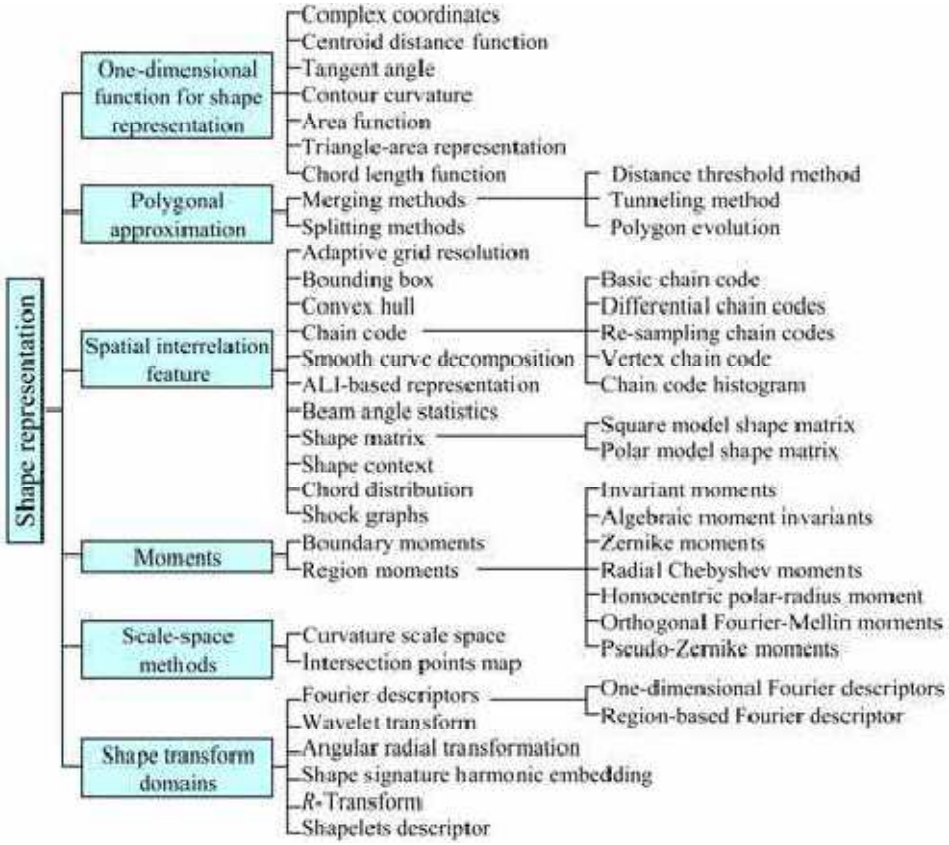

Figure: 3.3 An overview of shape description techniques

(a) Following feature are used in shaped based extraction:-

(1) *Circularity (Roundness)*

Circularity represents how a shape is similar to a circle. There are 3 defnitions and can be find by circularity ratio multiply with

$$\pi r^2 C.R.$$

. Circularity ratio is the ratio of the area of a shape to the area of a circle having the same perimeter:

$$C1 = \frac{A_S}{A_c} \qquad (2.24)$$

where As is the area of the shape and Ac is the area of the circle having the same perimeter as the shape. Assume the perimeter is O, so Ac = O/4π. Then C1 = 4π. As/O2. As 4π is a constant, so we have the third circularity ratio definition.

Circularity ratio is also called circle variance, and defined as:

$$Cva = \frac{\sigma_r}{\mu_r} \qquad (2.25)$$

where σR and μR are the mean and standard deviation of the radial distance from the centroid to the boundary points (xi; yi); i ∈ [0;N - 1]: They are the following formulae respectively:

$$\mu_R = \frac{1}{N}\sum_{i=1}^{N-1} d_i \; and \; \sigma_R = \sqrt{\frac{1}{N}\sum_{i=1}^{N-1}(d_i - \mu_R)^2} \qquad (2.26)$$

In order to extract shape based features we have used a software, Digimizer, which we have discussed in the coming section.

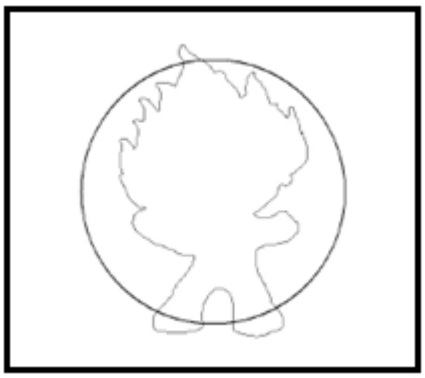

Figure: 3.4 Roundness feature

(2) *Area:*-Area feature tells the area within the selected region. It is find by area function (x(n),y(n))As boundary points change, the area S(n) of the triangle formed by triplet: (Pn, P_n+1, g) where g= (gx, gy) is centroid.

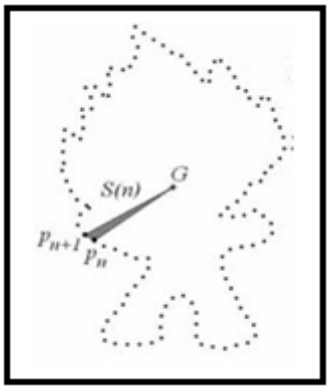

Figure: 3.5 Area feature

(3) *Parameter:-* Parameter feature tells the parameter within the selected region.

(b) Chord Length and Width Feature:-

Chord Length and width derived from reference point.For each boundary point P, Chord Length function CL. Shortest distance between P and another boundary pt P' subject to. PP'.At tangent vector at P. Chord Length function Invariant to translation and overcomes biased reference point problems (centroidbiased to boundary noise or deformations).Cord length and width is very sensitive to noise. Chord length can increase or decrease significantly even for smoothed boundaries.

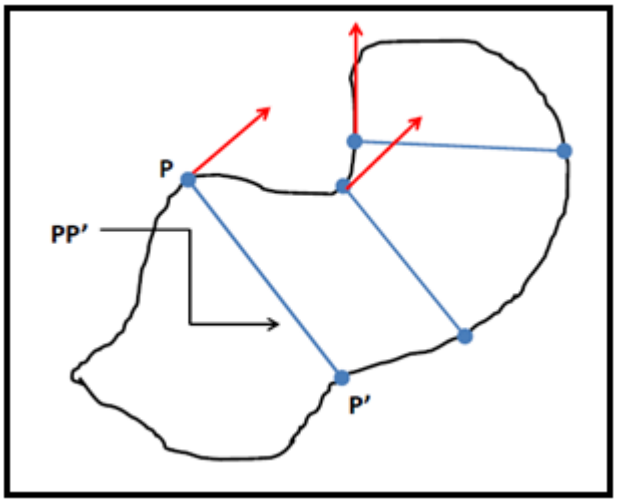

Figure: 3.6 Length and width Feature

3.4.1 Introduction to Digimizer Software

Digimizer is a free easy-to-use and flexible imageanalysis software package that allows precise manual measurements as well as automatic object detection with measurements of object characteristics.

Pictures may be X-rays, micrographs etc. Supported file formats are JPG, GIF, TIFF, BMP, PNG, WMF, EMF and DICOM files.

Image manipulation includes Resize, Cropping, Rotate, Flip, Zoom, Adjust contrast & brightness, Contrast auto fix, Stretch histogram, Background correction, Despeckle, Convert to grayscale, Invert, Negative, Sharpen.

Filters include Emboss, Arithmetic Mean Filter, Geometric Mean Filter, Harmonic Mean Filter, Median Filter, Maximum Filter, Minimum Filter, Midpoint Filter, Yp Mean Filter.

Digimizer allows to define unit of measurement, measure distances and lengths of line segments or paths, measure perimeters and areas, measure angles, locate middle of line segment, find center and calculate characteristics of circular objects, measurements on perpendicular lines, measure circles and rectangles, measure average intensity of objects, mark and count objects, fit lines.

Automatic image analysis includes Binarization, Options: overlay binary image, hide binary layer, Morphological operations: Dilate, Erode, Open, Close, Noise reduction, Analyze objects: object detection with measurement of perimeter and area.

The integrated statistics window displays statistics (n, mean, SD, minimum and maximum) of the measurements in the Measurements list.

Digimizer has been certified for Windows Vista, Windows 7 and Windows 8.

The different parts of the Digimizer application window are indicated in the following figure:

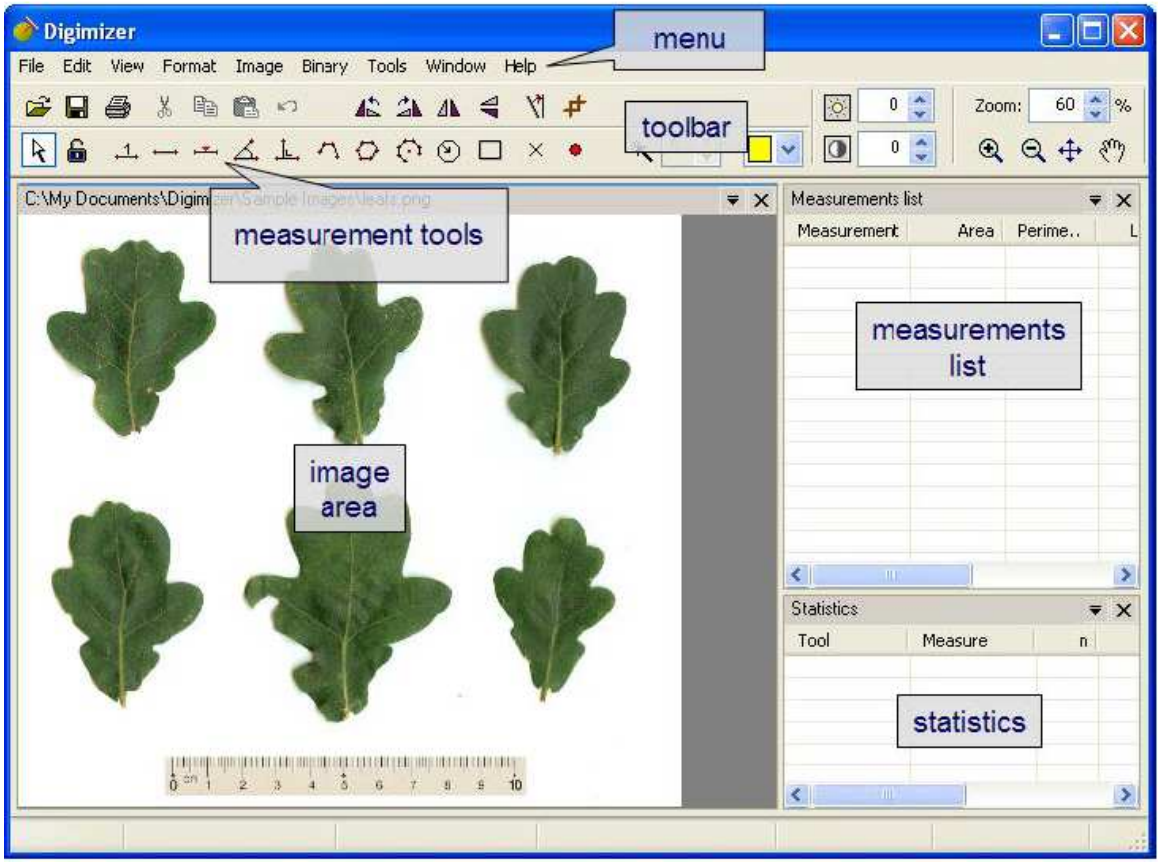

Figure: 3.7 Digimizer Tool Description

(a) Digimizer menu

In the menu you select the different commands and options.

(b) Toolbar

The toolbar contains buttons for the most common commands. Inactive buttons are displayed in gray. When you hover the mouse over an active button, a short description of the corresponding command is displayed in a small popup window. See p. 10 for a short description for all buttons on the toolbar.

(c) Measurement tools

The measurement tools toolbar contain the different tools for manual measurements in the image. For an overview of all measurements tools, see p. 11.

(d) Measurements list

The measurements list contains the different measurements performed in one or more images.

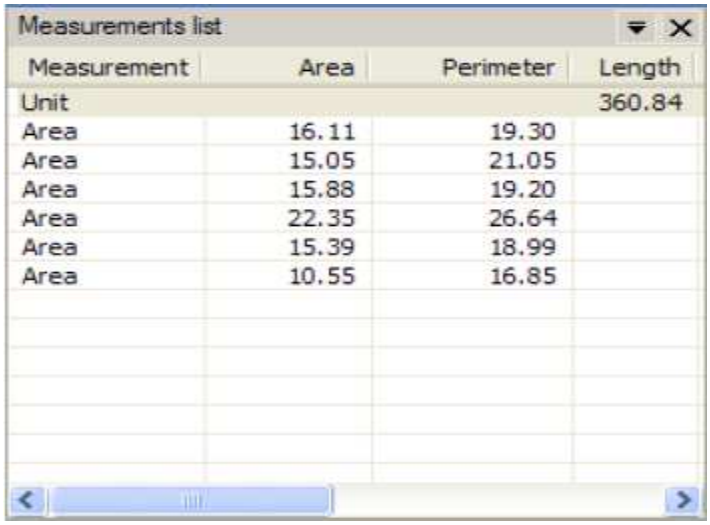

Figure: 3.8 Digimizer Measurement Parameter

(e) Statistics

The statistics window displays summary statistics for the different measurements that are displayed in the measurements list.

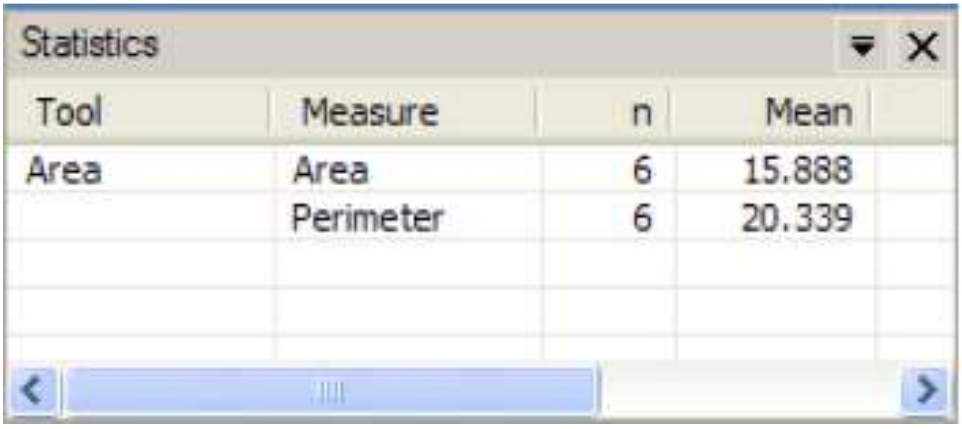

Figure:3.9 Digimizer Statistics Parameter

In Digimizer, you can save image in a special file format with file extension DGZ. The files in this format do not only contain the image in a compressed format (without quality loss), but also the measurements performed in this image.

3.5 Results

3.5.1 Texture Features:-

TABLE: 3.1 Texture Feature Extraction Results

Mask	Confusion Matrix	Accuracy (%)	S_B (%)	S_M (%)
3*3 Mask	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>32</td><td>1</td></tr><tr><td>M</td><td>21</td><td>4</td></tr></table>	62.0690	96.9	16
5*5 Mask	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>33</td><td>0</td></tr><tr><td>M</td><td>25</td><td>0</td></tr></table>	56.8966	1	0
7*7 Mask	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>31</td><td>2</td></tr><tr><td>M</td><td>21</td><td>4</td></tr></table>	60.344	93.9	16
9*9 Mask	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>32</td><td>0</td></tr><tr><td>M</td><td>25</td><td>0</td></tr></table>	56.1404	1	0

B: - Benign, M: - Malignant,S:-Sensitivity

3.5.2 Result of Shape Features and Shape along with Texture Feature :-

TABLE: 3.2 Shape feature and texture with shape feature results

	Confusion Matrix	Accuracy (%)	S_B (%)	S_M (%)
Shape features	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>27</td><td>6</td></tr><tr><td>M</td><td>9</td><td>16</td></tr></table>	74.13	81.81	64
Shape+Texture features	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>30</td><td>3</td></tr><tr><td>M</td><td>16</td><td>9</td></tr></table>	67.24	90.9	36

B: - Benign, M: - Malignant,S:-Sensitivity.

TABLE: 3.3 Dataset Information

Datasets	Benign	Malignant
Training Set	34	26
Testing set	33	25
Total	50	58

3.6 OUTCOME

3X3 Texture statistical feature extraction provides better accuracy than other Texture feature extraction. So we will take3X3 Texture statistical feature extraction for further evaluation.

CHAPTER 4

HYBRID CAD MODEL

4.1 Introduction

We have proposed a new Computer Aided Diagnosis model to increase the accuracy and lessen the false positive predictions. In this model we combine all the features of texture, shape and statistical data taken from MIAS database and Mammographic database. Already given, that there are too many instances in mammographic masses so reduction of database is required to merge with the texture and shape features of MIAS database.

4.2 Proposed Hybrid CAD Model

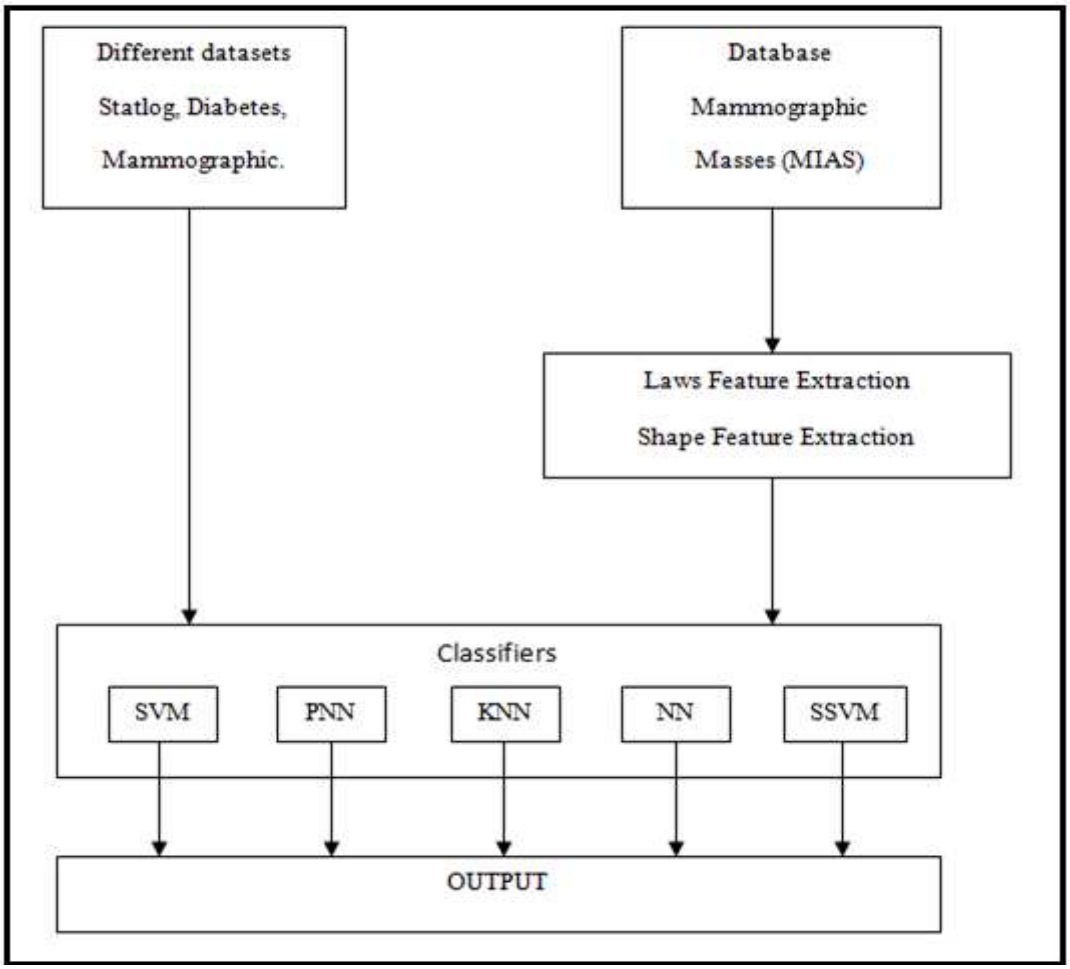

Figure: 4.1 Hybrid CAD Model

4.3 Steps to Follow

Following steps to be taken to perform the Hybrid CAD Model:-

1. Reduce The Dataset of Mammographic Masses to the size of MIAS Dataset by taking equal number of benign and malignant instances.
2. Combine the features of texture, shape and reduced statistical data.
3. Divide the new dataset into training and testing dataset.
4. Normalize the dataset and apply the SSVM classifier.
5. Find the accuracy and sensitivities.
6. Repeat all the above steps for texture and shape, and shape and reduced statistical data.

4.4 Advantages of Hybrid CAD System

A new and more efficient computer aided diagnosis classification method in digitized mammograms using Smooth Support Vector Machine (SSVM), which performs benign-malignant classification on statistical data based and image based region of interest (ROI) that contains mass, then combine to form a new diagnosis system . The major mammographic characteristics for diagnosis classification is texture, shape and some statistical data. SSVM exploits this important factor to classify it into benign or malignant. The statistical textural features used in characterizing the masses are mean, standard deviation, entropy, skewness and kurtosis and the shape features used in characterizing the masses are area, parameter, length, width, roundness. The main aim of this method and this system is to increase the effectiveness and efficiency of the classification process in an objective manner to reduce the numbers of false-positive of malignancies and benign prediction.

4.5 Results

TABLE: 4.1 Hybrid CAD Model Results

	Confusion Matrix	Accuracy (%)	S_B (%)	S_M (%)
Reduced Data (Mammographic Masses)	B: 29, 4 / M: 3, 22	87.93	87.87	88
Shape+Data	B: 27, 6 / M: 2, 23	86.2069	81.81	92
Texture+Shape	B: 30, 3 / M: 16, 9	67.2414	90.90	36
Texture+Shape +Data	B: 30, 3 / M: 1, 24	93.1034	90.90	96

B: - Benign, M: - Malignant, S:-Sensitivity

4.6 Outcome

Hybrid Model with all the feature of texture, shape and clinical feature gives the best accuracy result as compare to individual model as well other combination models.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

A proposed earlier hybrid model gives the best accuracy and effectiveness as compare to individual model viz clinical feature data, texture feature and shape feature. Also, hybrid model of all three features gives better accuracy as compare to combination of other models, as shown in results, by proving an accuracy of 93.1034%.

The future scope of work on this project can be that, to access both clinical and imaging data from the same patient, which will definitely boost the accuracy rate. Also,more texture feature extraction algorithm can be used, like GLCM that might increase the accuracy and effectiveness of all the proposed model given in this project.

REFERENCES

[1]. A comparative study on diabetes disease diagnosis using neural networks by Hasan Temurtas, Nejat Yumusak, Feyzullah Temurtas .

[2]. A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis by Mostafa Fathi Ganji, Mohammad Saniee Abadeh .

[3]. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier by Duygu Calistir, Esin Dog˘antekin.

[4]. Diabetes Data Analysis and Prediction Model Discovery Using Rapid Miner by Jianchao Han, Juan C. Rodriguze, Mohsen Beheshti.

[5]. A comparative study of computer-aided classification systems for focal hepatic lesions from B-mode ultrasound by Jitendra Virmani, Vinod Kumar, Naveen Kalra , and Niranjan Khandelwal.

[6]. Classification Of Diabetes Disease Using Support Vector Machine by V. Anuja Kumari, R.Chitra.

[7]. Computational Intelligence in Early Diabetes Diagnosis: A Review by Shankaracharya1, Devang Odedra1, Subir Samanta and Ambarish S. Vidyarth.

[8]. An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network by S.Priya and R.R.Rajalaxmi

[9]. Diagnosing Diabetes Type II Using a Soft Intelligent Binary Classification Model by Mehdi Khashei, Saeede Eftekhari, Jamshid Parvizian.

[10]. http://en.wikipedia.org/wiki/Diabetes_mellitus/

[11]. http://deckard.mc.duke.edu/breastcad.html

[12]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[13]. http://soe.cusat.ac.in/cs/workshop/coursedata/Lindo/NN-examples.pdf

[14]. http://en.wikipedia.org/wiki/Cardiovascular_disease

[15]. http://en.wikipedia.org/wiki/Breast_cancer

[16]. The prediction of breast cancer biopsy outcome using two CAD approaches tha both emphasize an intelligible decision processs by M.Elter,Schulz-Wendland and T.Wittenberg

[17]. Treatement of Missing Data Part-1 by David c.Howell

[18]. http://en.wikipedia.org/wiki/Support_vector_machine

[19]. http://www.academia.edu/8934493/Detection_of_Microcalcification_in_Mammograms_using_Soft_Computing_Techniques

[20]. http://en.wikipedia.org/wiki/Artificial_neural_network

[21]. http://www.mammoimage.org/databases/

[22]. http://peipa.essex.ac.uk/info/mias.html

[23]. http://marathon.csee.usf.edu/Mammography/Database.html

[24]. https://nf.nci.org.au/facilities/software/Matlab/toolbox/images/nlfilter.html

[25]. Intelligent Mammographic Database Management System for Computer aided breasr cancer by Isaac Adusei, Ognjen Kuljaca, Kwabena Agyepong Alcorn State University, System Research Institute

[26]. A Survey of Shape Feature Extraction TechniquesYang Mingqiang1,2, Kpalma Kidiyo1 and Ronsin Joseph1

[27]. http://in.mathworks.com/help/images/ref/nlfilter.html

[28]. SSVM: A Smooth Support Vector Machine forClassification Yuh-Jye Lee and O. L. Mangasarian Computer Sciences Department University of Wisconsin

[29]. A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis Santi Wulan Purnami, [1]Abdullah Embong, [1]Jasni Mohd Zain and [1]S.P. Rahayu

[30]. Enhancement of the image by using Histogram Modificationand High-pass Filtering Mask B.Suresh,U.Poojitha ,P.Vasanthi

[31]. Least Squares Support Vector Machine ClassifiersJ.A.K. SUYKENS and J. VANDEWALLE

APPENDIX

Inbuilt functions:-

1. **nlfilter**(A, [m n], fun) applies the function fun to each m-by-n sliding block of the grayscale image A. fun is a function that accepts an m-by-n matrix as input and returns a scalar result.
   c = fun(x)
   fun must be a function handle. See Parameterizing Functions, in the MATLAB Mathematics documentation, for information about how to provide additional parameters to the function fun.
   c is the output value for the center pixel in the m-by-n block x. nlfilter calls fun for each pixel in A. nlfilter zero-pads the m-by-n block at the edges, if necessary.
   B = nlfilter(A, 'indexed',...) processes A as an indexed image, padding with 1's if A is of class single or double and 0's if A is of class logical, uint8, or uint16.
2. **knnclassify** will be removed in a future release. Instead use fitcknn to fit a knn classification model and classify data using the predict function ofClassificationKNN object.
   Class = knnclassify(*Sample*, *Training*, *Group*)
3. Probabilistic neural networks (PNN) are a kind of radial basis network suitable for classification problems.net **= newpnn**(P,T,spread) takes two or three arguments.
4. **Newff:-** Create feed-forward backpropagation network It Feed-forward networks consist of Nl layers using the dotprod weight function, netsum net input function, and the specified transfer function.The first layer has weights coming from the input. Each subsequent layer has a weight coming from the previous layer. All layers have biases. The last layer is the network output.Each layer's weights and biases are initialized with initnw. Adaption is done with trains, which updates weights with the specified learning function. Training is done with the specified training function. Performance is measured according to the specified performance function.
5. **sim** :-It is used to Simulate neural network.
6. **ssvm_train**:-It is used to train the ssvm network.
7. **svm_train**:-It is used to train the svm network.