Skin Disease Prediction Using Neural Network Algorithm

Dr. S. Vijayarani[1], Ms. E. Suganya[2], Ms. S.Sowmiya[3]
Assistant Professor[1], Ph.D Research Scholar[2], PG Student[3]
Department of Computer Science
Bharathiar University

Abstract: Data mining is the process of analyzing the existing data and extracting useful information out of it. Sometimes data mining used to predict the future based on the existing data. It also finds the relation between the existing data and based on the relation it predicts the outcome of the remaining data. There are several methodologies used for these problems like classification, clustering, regression, rule generation, etc.Classification is a type of data mining algorithm used to predict class labels and classify the data to a particular class based on training data sample and then is used to classify the new test data sets.The main aim of the research work is to predict the six types of skin cancer diseases such as Melonoma, Basal cell carcinoma, Benign Keratosis-Like Lesions, Vascular Lesions, Dermatofibroma, andAkiec using data mining classification algorithms which are Random Forest, Recursive Partitioning, SVM-Linear, SVM-Kernel are used for classification. In addition to this, the research work neural network was proposed for classification and get better accuracy.

Keywords: Classification, Random Forest, Recursive Partitioning, SVM-Linear, SVM-Kernel, Neural Network

1. INTRODUCTION

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is used to mine or extractknowledge from large quantities of data. Data mining techniques are used to implement and solve the different types of research problems. The research related areas in data mining are text mining, web mining, image mining, sequential pattern mining, spatial mining, medical mining, multimedia mining, structure mining and graph mining. Data mining is used to identify valid novel, potentially useful, and understandable correlations and patterns from the existing data. Finding useful patterns in data is known by different names in different communities.It is an extension of traditional data analysis and statistical approaches in that it incorporates analysis techniques drawn from a range of disciplines including numerical analysis, pattern matching and areas of artificial intelligence such as machine learning, neural networks and genetic algorithms. While many data mining tasks follow a traditional, hypothesis-driven data analysis approach, it is a common place to employ an opportunistic, data driven approach that encourages the pattern detection algorithm to find useful trends, patterns and relationships. There are several important techniques are available in data mining which includes association classification, clustering, prediction, sequential patterns, time series analysis, privacy preservation and etc.

Classification is a typical data mining technique which is based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification applications include, image and pattern recognition, medical diagnosis, loan approval detecting faults in industry applications classifying financial market trends. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics. For example, to apply classification in the application that, given all records of employees who left the company, predict who will probably leave the company in a future period. In this case, to divide the records of employees into two groups that named leave and stay. After that, use the data mining software to classify the employees into separate groups.Classification means maps data in to predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. The classification algorithm requires that the classes be defined based on data attribute values. They often describe these classes by looking at the characteristics of data already known to those classes. It's a type of classification where an input pattern is classified into one of several classes based on its similarity to these predefined classes.

The remaining sections are represented as follows: section 2 discussed about the literature review, section 3 described about methodology which includes the existing and proposed classification algorithms.The experiments and results are shown in section 4 and the conclusion in given in section 5.

2. REVIEW OF LITERATURE

Divya Jain et al. [1] have reviewed the current feature selection approaches and classification systems for effective disease prediction. The performance metrics used in medical diagnostic systems to measure the performance of the classification methods are also discussed. The authors evaluated k-nearest neighbour algorithm, Naïve Bayes classifier.

Vikas Chaurasia et al. [2] have evaluated classification techniques based on the selected classifier algorithm. Sequential Minimal Optimization (SMO), IBK and BF Tree are used. SMO shows the results with Breast Cancer disease of patient records. Therefore SMO classifier is suggested for diagnosis of Breast Cancer disease classification to get better results.

Meherwar Fatima et al. [3] surveyed different machine learning techniques for the diagnosis of different diseases such as heart disease, diabetes disease, liver disease, dengue and hepatitis disease. Many algorithms have shown good results because they identify the attribute accurately. The authors implemented the Naïve Baye's, j48, CART, ID3, and Random Forest classification algorithms.

Subhash Chandra et al. [4] discussed a study of different data mining techniques that can be employed in robotic heart disease prediction systems. The symbolic Fuzzy K-NN classifier can be tested with the unstructured data available in health care industry data. Collection of number of records to provide better accuracy to the system in predicting and diagnosing the patients of heart disease.

Dr.S.Vijayarani et al. [6] have evaluated the Naïve Bayes and Support Vector Machine classification algorithms. Classification process is used to classify four types of kidney diseases. Comparison of Support Vector Machine (SVM) and Naïve Bayes classification algorithms is done based on the performance. From the results, the authors concluded that the SVM increased the classification performance. Hence it was considered as best classifier.

Dr.S.Vijayarani et al.[7] analyzed classification algorithms namely Naïve Bayes and Support Vector Machine (SVM) for liver disease prediction. The comparisons of these algorithms are done. From the experimental results, the authors concluded that the SVM classifier is considered as the best algorithm because of its highest classification accuracy.

Dhyan Chandra Yadavet et al. [8] have compared the Bayesian and Lazy classifier algorithms for classifying consumer claim data set. The Bayesian Algorithm includes two techniques namely Bayes Net, Naïve Bayes and the Lazy algorithms includes IBl (Instance Based Learning), IBK (K-Nearest Neighbour) and KStar techniques. They concluded that two algorithms such as IBI and KStar gives better results than other algorithms

K.Arutchelvan et al. [9] developed a system to provide earlier warning to the users. It predicted three specific cancer risks. Specifically, the cancer prediction system estimates the risk of the breast, skin and lung cancers by examining a number of users provided genetic and non-genetic factors. People can easily check their risk and take appropriate action based on their risk status.

3. METHODOLOGY

Methodology is the process of developing a system through successive phases in an orderly way. Here classification methods are used to predict the skin cancer disease. Classification is one of the data mining techniques. It is otherwise known as supervised learning that assigns categories or continuous values to a collection of data in order to produce more accurate predictions and analysis. Mostly, classification is used to classify each item in a data set into one of a predefined set of classes or groups.

3.1 EXISTING ALGORITHMS

Random Forest

Random forest is a machine learning technique. It is an ensemble classifier because it consists of many decision trees and output the classes by individual trees. The collection of decision trees is to construct with controlled variation for random selection of features. It performs classification, regression and dimensionality reduction. Each attribute is classified by the tree vote according to the vote the random forest choose the class which having the most votes. Random forest is a collection of unpruned decision trees. It is mostly used for very large training datasets and a very large input variable. It consists of many decision trees and output the class that is the mode of the class output by individual trees. It produces a highly accurate classifier for many datasets. It generates an internal unbiased estimate of the generalization error as the forest building progresses and also includes a good method for estimating missing data and maintains accuracy when a large proportion of the data are missing. The advantages of random forests are: It provides an experimental way to detect variable interactions, It can balance error in the class population of unbalanced data sets. It computes proximities between cases, useful for clustering, detecting outliers, and (by scaling) visualizing the data. Its learning is fast.

Recursive Partitioning

A recursive partitioning algorithm is used for classification by building both decision trees and can also be used to generate regression trees. To generate the structure of a tree, by using the decision tree rules for predicting a categorical (classification tree) or continuous (regression tree) outcome. The recursive partitioning programs build classification or regression models of a very general structure that is a two-stage procedure; the resulting models can be as binary trees.

The tree is built by the following process:

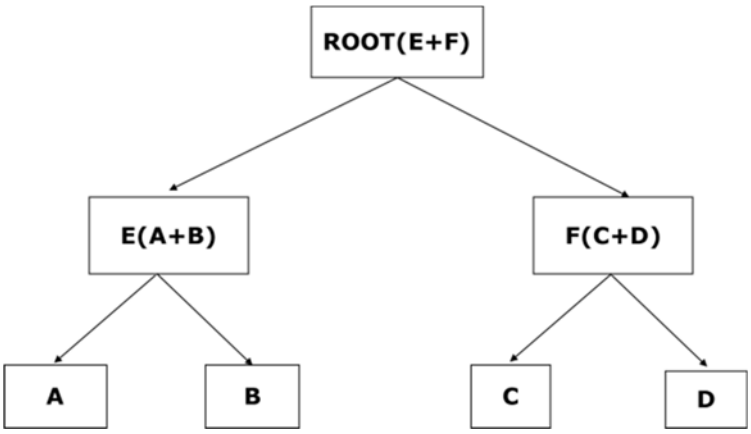


Figure 3.2 Example of Recursive Partitioning Tree

Support Vector Machines (kernels)

In SVM, it is easy to have a linear hyper-plane to classify two classes. SVM has a technique called the kernel which is useful for non-linear hyperplane classification. These are functions which transform low dimensional input space in to a higher dimensional space i.e. it converts non-classified problem to classified problem, these functions are called kernels. It is mostly useful in non-linear separation problem. There are some kernels which are used for the transformation of low dimension space to high dimension space; they are radial, linear and polynomial. In this work two types are used of SVM kernels namely SVM Radial Kernel method and SVM Linear Kernel method.

Linear kernel SVM

The dot-product is used which is called kernel and it will be written as:

$$K(x, x_i) = \text{sum}(x * x_i)$$

Here k is the kernel that defines the similarity or a distance measure between new data and the support vectors. The dot product is the similarity measure used for linear kernel because the distance is a linear combination of the inputs.

Radial Kernel SVM

The Radial kernel is more complex to linear kernel. For example:

$$K(x, x_i) = \exp(-\text{gamma} * \text{sum}((x - x_i\text{^}2))$$

Where gamma is a parameter that must be used in support vector machine learning algorithm. Gamma 0.1 is a good default value, where gamma is between 0 to1. The radial kernel can create complex regions within the feature space and transform low dimensional space to high dimensional space.

3.2 PROPOSED ALGORITHM

Neural Network for Skin Disease Prediction (NNSDP)

Neural networks are a new paradigm in computing. It involves developing mathematical structures with the ability to learn. The methods are the result of academic attempts to model the nervous system learning. Neural networks have the remarkable ability to derive meaning from Complicated or imprecise data. It can be used to extract patterns and detect trends that are too complex to be noticed by either humans are or other computer techniques. Neural networks are better at identifying patterns or trends in data; they are well suited for prediction or forecasting needs. Neural networks are used as a set of processing elements analogous to neurons in the brain. These processing elements are interconnected in a network that can then identify patterns in data once it is exposed to the data, i.e., the networks learns from experience just as people do. Neural networks often referred to as artificial neural networks (ANN). The NN is actually an information processing system. It consists of a graph representing the processing system as well as various algorithms that access that graphically.

As with the human brain, the NN consists of many connected processing elements. The NN is structured as a directed graph with many nodes (processing elements) and arcs (interconnections) between them. The nodes in the graph are like individual neurons, while the arcs are their interconnections. Each of these processing elements functions independently from the others and uses only local data (input and output to the node) to direct its processing. This feature facilitates the use of NNs in a distributed and/or parallel environment. The NN approach, like decision trees, requires that a graphical structure to be built to represent the model and then that the structure be applied to the data can be viewed as a directed graph with source (input), sink (output), and internal (hidden) nodes.

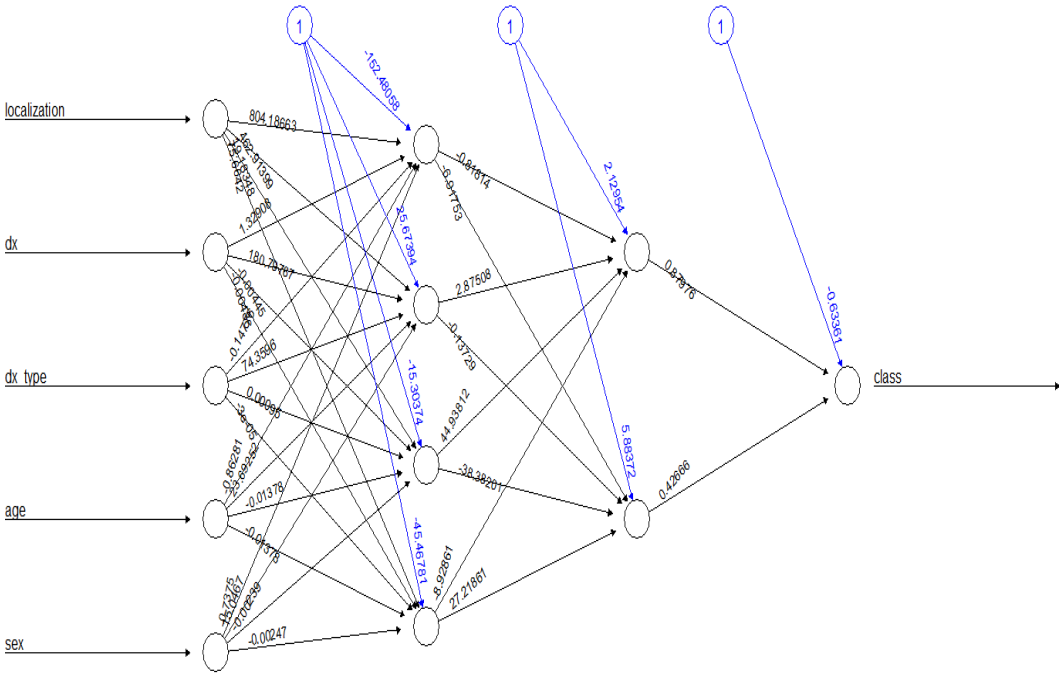
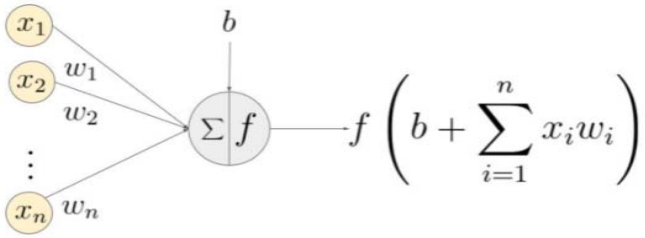


Figure3.6 Neural Network

First, we determine the basic structure of the graph. There are five important attributes, assume that there are five input nodes. Classify into one input layers, two hidden layers and one output layer. The number of hidden layers in the NN is not easy to determine. In most cases, one or two is enough. In this figure, assume that there is two hidden layers and thus a total of four layers. There are two nodes in the hidden layer. Each node is labeled with a function that indicates its effect on the data coming into that node.

$$f\left(b+\sum_{i=1}^{n} x_i w_i\right)$$



In the above figure3.6 *(x1,x2,x3,x4,x5)* is the input signal vector that gets multiplied with the weights(*w1,w2,…,w20*). This is followed by accumulation (i.e. summation + addition of bias *b*). Finally, an activation function *f* is applied to this sum. In the above figure (*x1,x2,x3,x4,x5*) is the input signal vector that gets multiplied with the weights(*w1,w2,…,w20)* This is followed by accumulation (i.e. summation + addition of bias *b*). Finally, an activation function *f* is applied to this sum.

Pseudo code for Neural Network

1. Initialize NN model
2. Initialize training algorithm parameters
3. Assign random weights to all linkages to start the algorithm
4. Find the activation rate of hidden nodes
5. Using the activation rate of hidden nodes and linkages to output find the activation rate of output nodes
6. Find the error rate at the output node and recalibrate all the linkages between hidden nodes and output nodes
7. Using the weights and error found at output node, cascade down the error to hidden nodes
8. Recalibrate the weights between the hidden node and the input nodes
9. Repeat the process till the convergence criterion is met
10. Using the final linkage weights scores the activation rate of output nodes

4. EXPERIMENTAL RESULTS

This experiment is conducted using R-Studio on the system with an AMD PRO A4-3350B processor running at 2.00 GHz, 4 GB RAM, 32-bitWindows 8. A performance test is evaluated based on search time.

4.2 DATASET DESCRIPTION

In order to perform the experiments, skin dataset is collected from Kaggle.com which contains 10015 instances and 8 attributes such as lesion_id, image_id, dx, dx_type, localization and class. Dataset contains the six type of skin cancer diseases (i.e.) Melonoma, Basal cell carcinoma, Benign Keratosis-Like Lesions, Vascular Lesions,Dermatofibroma and Akiec. The main objective of this work is to identify the following (i) skin cancer type classification, (ii) No. of persons affected with melanoma, (iii) No. of persons affected with Basal cell Cancer, (iv) No. of persons affected with Benign Keratosis-like lesions, (v) No. of persons affected with Melanocytic Nevi, (vi) No. of persons affected with Vascular Lesions, (vii) No. of persons affected with Dermatofibroma, (viii) No. of persons affected with Akiec, (ix) Localization based skin cancer classification.

4.3 CLASSIFICATION ALGORITHMS

4.3.1 Algorithm Accuracy

Table 4.1 Classification Accuracy

Algorithms	Correctly Classified Instances (%)	Incorrectly Classified Instances (%)
Random forest	94.43	5.57
Recursive Partitioning	94.2	5.8
SVM Radial kernel	85.1	14.9
SVM-Linear	89.3	10.7
NNSDP	98.3	1.7

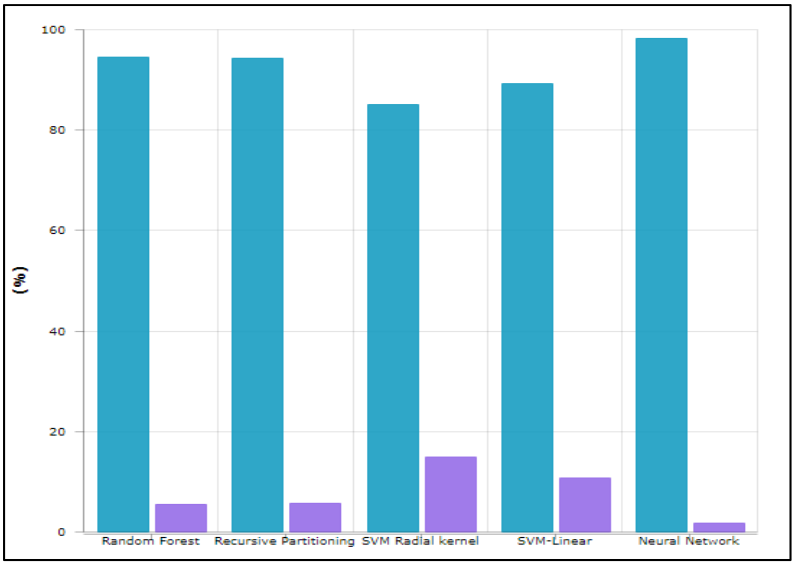


Figure 4.1 Classification Accuracy

Table 4.1 and figure 4.1 shows the accuracy measures of classification algorithms. The performance factors of classification accuracy are correctly classified instances, incorrectly classified instances.

4.4 Age wise Report

4.4.1 Benign Keratosis-Like Lesions based Age wise Report

Table 4.2 Benign Keratosis-Like Lesions based Age wise Report

Age	Gender		
	Male	Female	Total
5-25	9	14	23
25-45	75	118	193
45-65	311	168	479
65-85	294	124	418

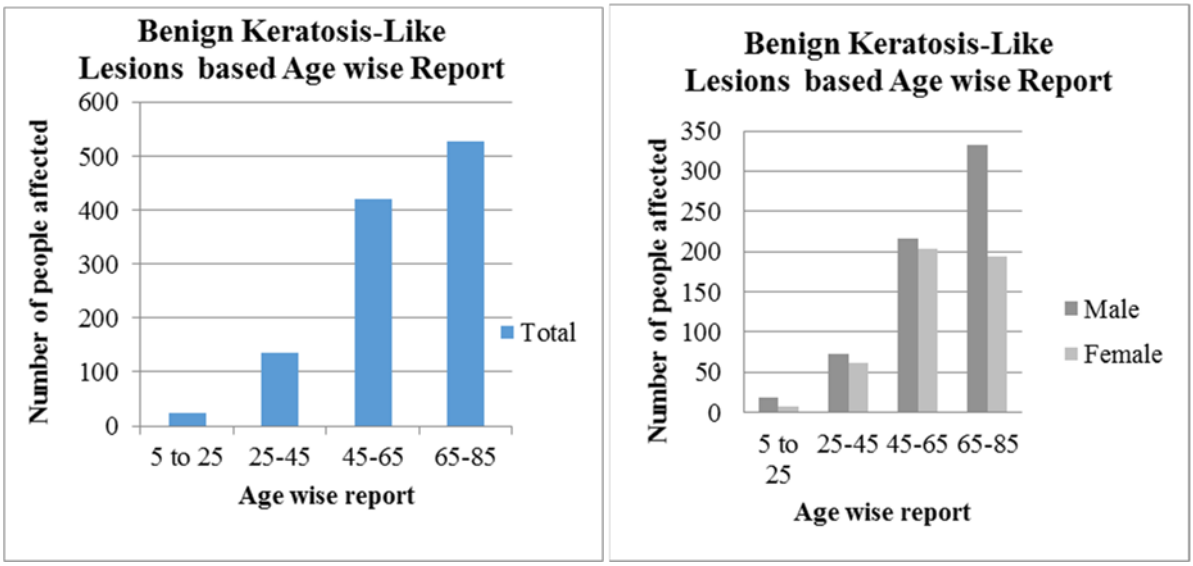


Figure 4.2 Benign Keratosis-Like Lesions based Age wise Report

Table 4.3 shows the age wise report of people who affected by the skin cancer based on dx attribute. From this analysis many people who have age between 65-85 are affected by the BKL skin cancer, and mostly male peoples are affected by this kind of disease.

4.4.2 Melonoma based Age wise Report

Table 4.3 Melonoma based Age wise Report

Age	Gender		
	Male	Female	Total
5-25	3	1	14
25-45	31	27	58
45-65	90	76	166
65-85	193	93	286

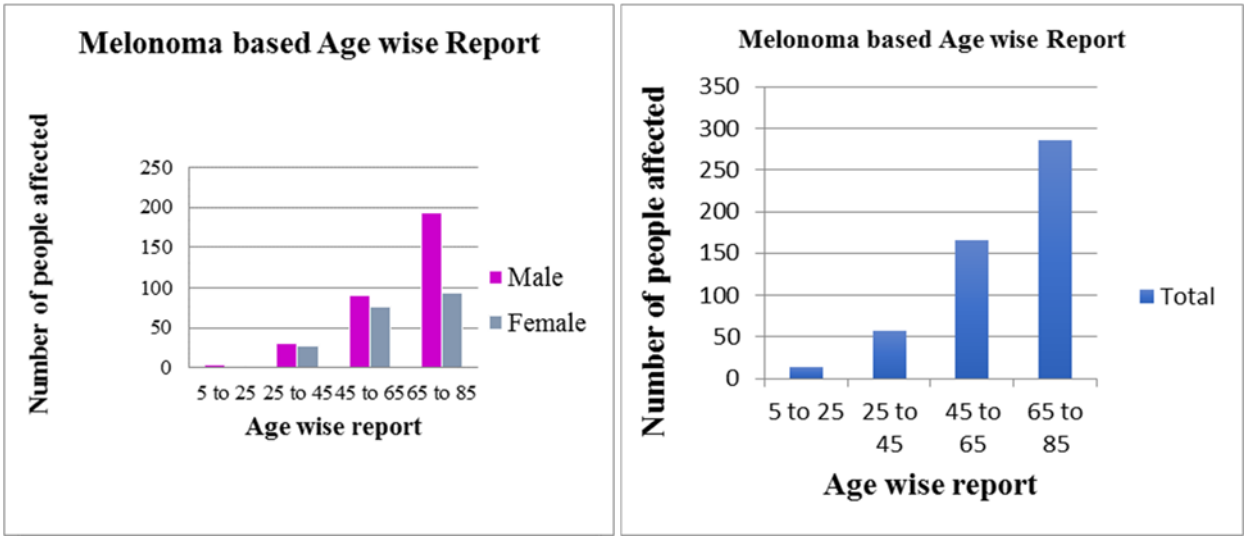


Figure 4.3 Melonoma based Age wise Report

Table 4.3 and figure 4.3 shows the age wise report of people who affected by the skin cancer based on dx attribute. From this analysis many people who have age between 65-85 are affected by the Melonoma skin cancer, and mostly male peoples are affected by this kind of disease.

4.4.3 Basal Cell Cancer based Age wise Report

Table 4.4 Basal Cell Cancer based Age wise Report

Age	Gender		
	Male	Female	Total
5-25	13	9	22
25-45	9	23	32
45-65	21	26	47
65-85	26	15	41

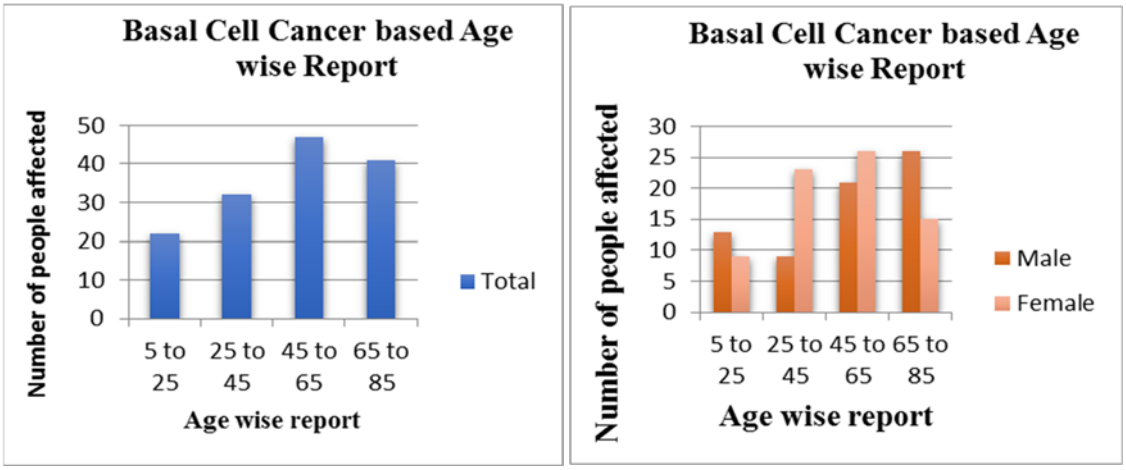


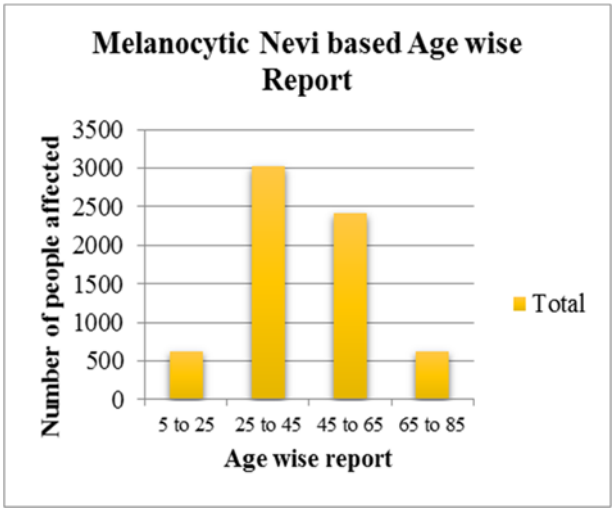
Figure 4.4 Basal Cell Cancer based Age wise Report

Table 4.4 and figure 4.4 shows the age wise report of people who affected by the skin cancer based on dx attribute. From this analysis many people who have age between 45-65 are affected by the Basal cell skin cancer, and mostly female peoples are affected by this kind of disease.

4.4.4 Melanocytic Nevi based Age wise Report

Table 4.5 Melanocytic Nevi based Age wise Report

Age	Gender		
	Male	Female	Total
5-25	287	330	617
25-45	1388	1636	3024
45-65	1317	1102	2419
65-85	453	165	618



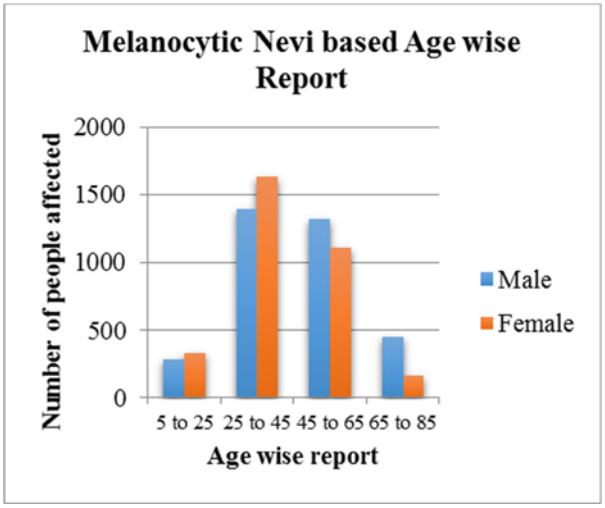


Table 4.5 Melanocytic Nevi based Age wise Report

Table 4.5 and figure 4.5 shows the age wise report of people who affected by the skin cancer based on dx attribute. From this analysis many people who have age between 25-45 are affected by the **Melanocytic Nevi** skin cancer, and mostly male peoples are affected by this kind of disease.

4.4.5 Vascular Lesions based Age wise Report

Table 4.6 Vascular Lesionsbased Age wise Report

Age	Gender		
	Male	Female	Total
5-25	6	2	10
25-45	16	23	39
45-65	34	24	58
65-85	13	3	16

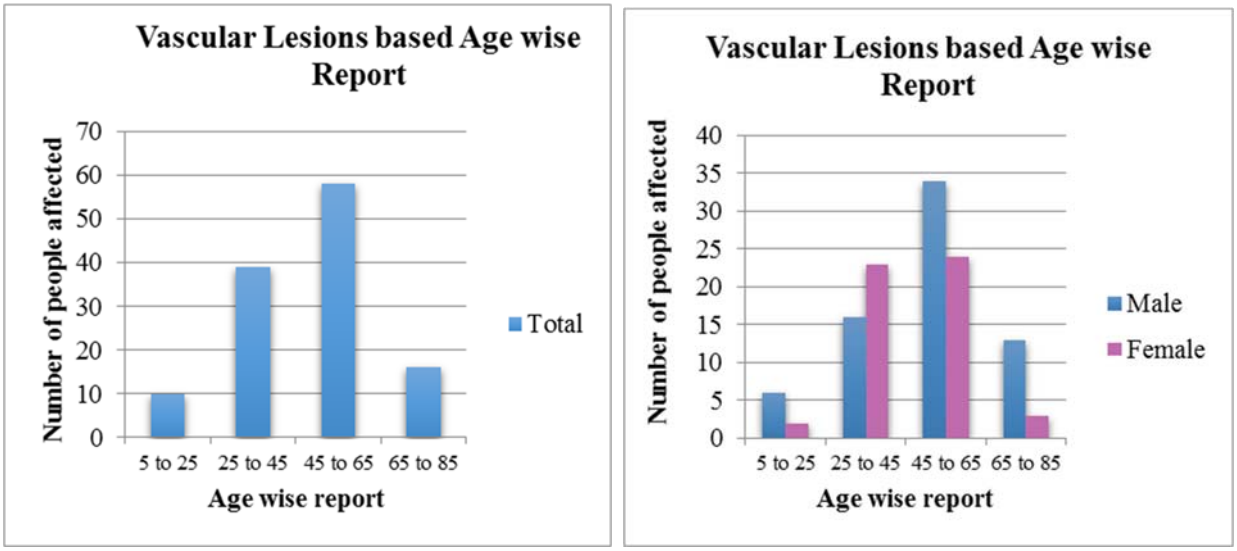


Figure 4.6 Vascular Lesionsbased Age wise Report

Table 4.6 and figure 4.6 shows the age wise report of people who affected by the skin cancer based on dx attribute. From this analysis many people who have age between 45-65 are affected by the Vascular Lesions skin cancer, and mostly male peoples are affected by this kind of disease.

4.4.6 Dermatofibroma based Age wise Report

Table 4.7 Dermatofibroma based Age wise Report

Age	Gender		
	Male	Female	Total
5-25	-	-	-
25-45	17	3	20
45-65	90	52	142
65-85	114	51	165

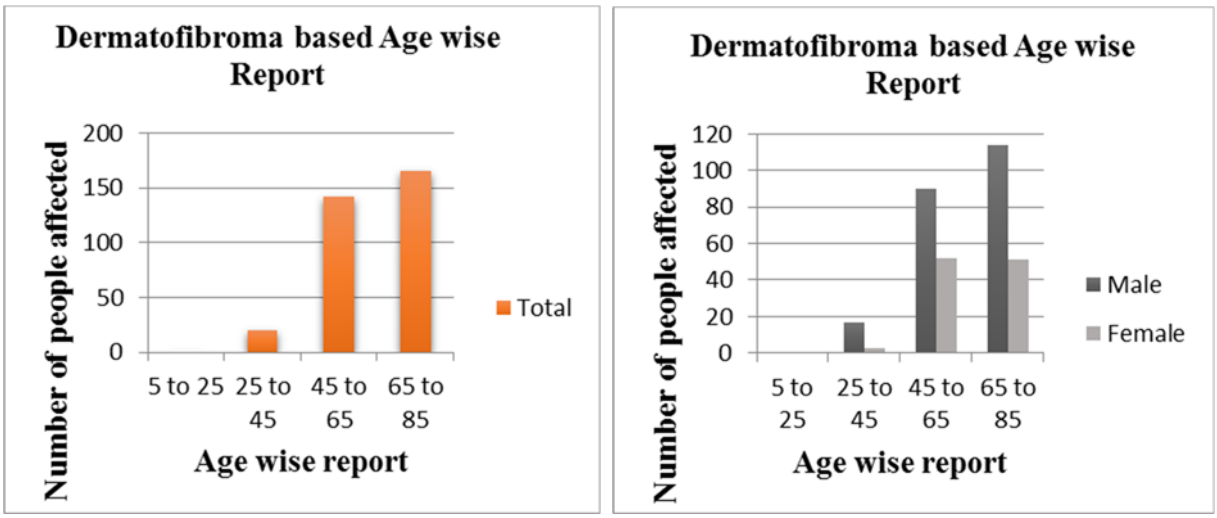


Figure 4.7 Dermatofibroma based Age wise Report

Table 4.7 and figure 4.7 shows the age wise report of people who affected by the skin cancer based on dx attribute. From this analysis many people who have age between 65-85 are affected by the Dermatofibroma skin cancer, and mostly male peoples are affected by this kind of disease.

4.4.7 Akiec based Agewise Report

Table 4.8 Akiec based Agewise Report

Age	Gender		
	Male	Female	Total
5-25	18	7	25
25-45	73	62	135
45-65	217	204	421
65-85	333	194	527

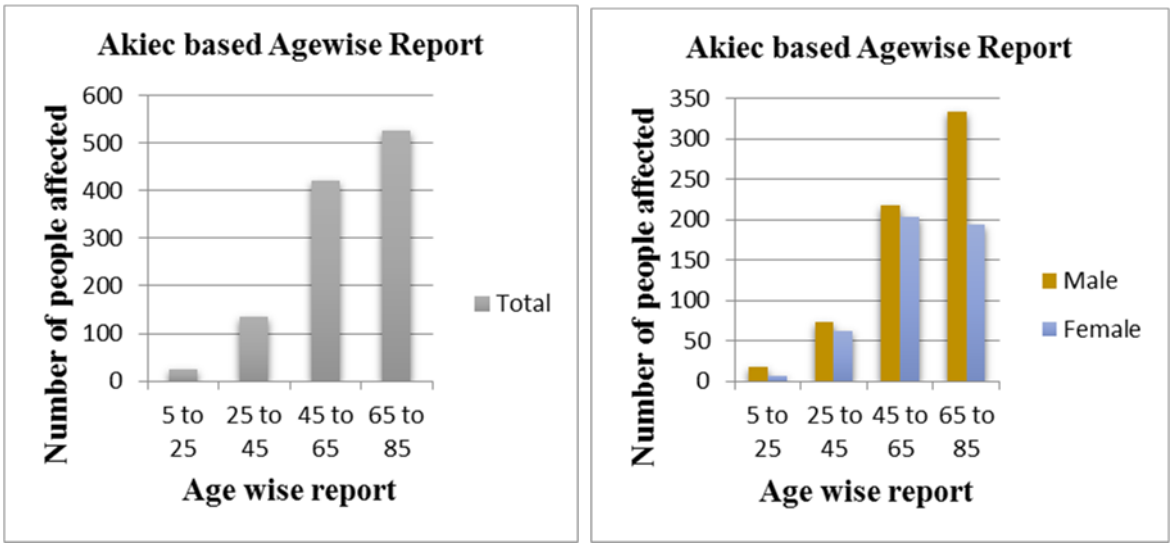


Figure 4.8 Akiec based Agewise Report

Table 4.8 and figure 4.8 shows the age wise report of people who affected by the skin cancer based on dx attribute. From this analysis many people who have age between 65-85 are affected by the Dermatofibroma skin cancer, and mostly male peoples are affected by this kind of disease.

4.5 Benign Keratosis-Like Lesions – Localization

Table 4.9 BKL-Localization based Agewise Report

Localization	Age	Gender		
		Male	Female	Total
Scalp	5-25	1	-	1
	25-45	1	-	1
	45-65	9	-	9
	65-85	21	3	24
Ear	5-25	-	-	-
	25-45	-	-	-
	45-65	-	2	2
	65-85	4	-	4
Face	5-25	2	-	2
	25-45	23	24	47
	45-65	60	72	132
	65-85	74	66	140
Back	5-25	-	3	3
	25-45	11	5	16

	45-65	48	14	62
	65-85	88	33	121
Trunk	5-25	-	-	-
	25-45	3	5	8
	45-65	14	7	21
	65-85	24	20	44
Chest	5-25	-	-	-
	25-45	36	-	36
	45-65	-	-	-
	65-85	23	35	85
Upper extremity	5-25	-	-	-
	25-45	3	3	6
	45-65	4	-	4
	65-85	-	8	8
Unknown	5-25	-	-	-
	25-45	4	3	7
	45-65	9	8	17
	65-85	5	5	10
Abdomen	5-25	-	-	-
	25-45	74	66	140
	45-65	-	-	-
	65-85	4	3	7
Lower extremity	5-25	9	8	17
	25-45	-	-	-
	45-65	88	33	121
	65-85	3	5	8
Genital	5-25	2	5	7
	25-45	9	8	17
	45-65	74	66	140
	65-85	50	70	120
Neck	5-25	9	-	9
	25-45	67	43	110
	45-65	9	8	17
	65-85	-	-	-
Hand	5-25	8	10	18
	25-45	74	66	140
	45-65	45	12	57
	65-85	88	33	121
Foot	5-25	-	-	-
	25-45	10	12	22
	45-65	9	8	17
	65-85	4	3	7

Table 4.9 shows the age wise report of people who affected by the skin cancer based on localization attribute. Results Observed by the performance of the algorithms. From this analysis many people affected by the Benign Keratosis-Like Lesions disease on their face and hand. Mostly many people who have age between 65-85 are affected by the Benign Keratosis-Like Lesions skin cancer on their Face and Hand, and mostly male peoples are affected by this kind of disease.

4.6 Melonoma – Localization

Table 4.10 Mel-Localization based Age wise Report

Localization	Age	Gender		
		Male	Female	Total
Scalp	5-25	-	1	1
	25-45	-	-	-
	45-65	-	1	1
	65-85	12	-	12
Ear	5-25	-	-	-
	25-45	-	2	-
	45-65	5	1	6
	65-85	6	3	9
Face	5-25	-	1	1
	25-45	3	3	6
	45-65	33	14	47
	65-85	30	20	50
Back	5-25	6	3	9
	25-45	36	17	53
	45-65	114	48	162
	65-85	69	31	100
Trunk	5-25	-	-	-
	25-45	4	3	7
	45-65	18	3	21
	65-85	17	2	19
Chest	5-25	10	12	22
	25-45	88	33	121
	45-65	9	8	17
	65-85	9	7	16
Upper extremity	5-25	74	66	140
	25-45	-	-	-
	45-65	10	12	22
	65-85	9	8	17
Unknown	5-25	-	-	-
	25-45	88	33	121
	45-65	80	2	82
	65-85	74	66	140
Abdomen	5-25	-	-	-
	25-45	4	3	7
	45-65	9	8	17

	65-85	10	12	22
Lower extremity	5-25	74	66	140
	25-45	7	12	19
	45-65	88	33	121
	65-85	9	8	17
Genital	5-25	-	-	-
	25-45	17	15	32
	45-65	88	33	121
	65-85	-	-	-
Neck	5-25	9	8	17
	25-45	4	3	7
	45-65	10	12	22
	65-85	74	66	140
Hand	5-25	-	-	-
	25-45	74	66	140
	45-65	88	33	121
	65-85	9	8	17
Foot	5-25	9	7	16
	25-45	88	33	121
	45-65	74	66	140
	65-85	76	21	97

Table 4.10 shows the age wise report of people who affected by the skin cancer based on localization attribute. Results Observed by the performance of the algorithms. From this analysis many people affected by the Melonoma disease on their face, chest, foot, trunk and hand. And mostly many people who have age between 5-85 and 25-45 are affected by the Melonoma skin cancer, and mostly male peoples are affected by this kind of disease.

4.11 Basal cell carcinoma – Localization

Table 4.11 BCC-Localization based Agewise Report

Localization	Age	Gender		
		Male	Female	Total
Scalp	5-25	-	-	-
	25-45	-	-	-
	45-65	3	3	6
	65-85	9	4	13
Ear	5-25	-	-	-
	25-45	-	-	-
	45-65	-	-	-
	65-85	-	-	-
Face	5-25	-	1	1
	25-45	4	2	6
	45-65	15	10	25
	65-85	38	31	69
Back	5-25	3	-	3
	25-45	-	12	12
	45-65	31	32	63

	65-85	89	25	94
Trunk	5-25	-	-	-
	25-45	-	2	2
	45-65	6	-	6
	65-85	-	-	-
Chest	5-25	88	33	121
	25-45	74	66	140
	45-65	-	-	-
	65-85	9	8	17
Upper extremity	5-25	-	-	-
	25-45	20	40	60
	45-65	74	66	140
	65-85	9	6	15
Unknown	5-25	2	6	8
	25-45	9	8	17
	45-65	88	33	121
	65-85	-	-	-
Abdomen	5-25	74	66	140
	25-45	-	-	-
	45-65	54	12	66
	65-85	9	8	17
Lower extremity	5-25	-	-	-
	25-45	74	66	140
	45-65	11	9	20
	65-85	4	3	7
Genital	5-25	9	8	17
	25-45	88	33	121
	45-65	77	1	78
	65-85	74	66	140
Neck	5-25	-	-	-
	25-45	9	8	17
	45-65	-	-	-
	65-85	74	66	140
Hand	5-25	88	33	121
	25-45	8	9	17
	45-65	6	12	18
	65-85	9	8	17
Foot	5-25	18	3	21
	25-45	74	66	140
	45-65	23	22	55
	65-85	88	33	121

Table 4.11 shows the age wise report of people who affected by the skin cancer based on localization attribute. Results Observed by the performance of the algorithms. From this analysis many people affected by the Basal cell skin cancer disease on their face, foot and hand. And mostly many people who have age between 5-25 and 45-65 are affected by the Basal cell skin cancer, and mostly female peoples are affected by this kind of disease.

4.7 Melanocytic Nevi – Localization

Table 4.12 Melanocytic Nevi-Localization based Age wise Report

Localization	Age	Gender		
		Male	Female	Total
Scalp	5-25	3	3	6
	25-45	9	11	20
	45-65	9	5	14
	65-85	4	-	4
Ear	5-25	-	3	3
	25-45	3	3	6
	45-65	1	14	15
	65-85	2	2	4
Face	5-25	11	13	24
	25-45	13	25	38
	45-65	13	13	26
	65-85	10	-	10
Back	5-25	60	62	122
	25-45	330	312	642
	45-65	298	196	494
	65-85	133	34	167
Trunk	5-25	25	26	51
	25-45	254	327	584
	45-65	288	213	501
	65-85	86	19	105
Chest	5-25	-	-	-
	25-45	74	66	140
	45-65	86	19	105
	65-85	-	-	-
Upper extremity	5-25	4	3	7
	25-45	88	33	121
	45-65	254	327	584
	65-85	86	19	105
Unknown	5-25	9	8	17
	25-45	44	22	66
	45-65	74	66	140
	65-85	86	19	105
Abdomen	5-25	254	327	584
	25-45	-	-	-
	45-65	12	15	27
	65-85	9	8	17

Lower extremity	5-25	17	4	21
	25-45	32	12	44
	45-65	74	66	140
	65-85	-	-	-
Genital	5-25	86	19	105
	25-45	88	33	121
	45-65	43	13	56
	65-85	-	-	-
Neck	5-25	74	66	140
	25-45	254	327	584
	45-65	-	17	17
	65-85	9	8	17
Hand	5-25	23	22	45
	25-45	88	33	121
	45-65	-	17	17
	65-85	74	66	140
Foot	5-25	12	21	33
	25-45	254	327	584
	45-65	55	-	55
	65-85	88	33	121

Table 4.12 shows the age wise report of people who affected by the skin cancer based on localization attribute. Results Observed by the performance of the algorithms. From this analysis many people affected by the Melanocytic Nevi skin cancer disease on their neck, foot, upper extremity and hand. And mostly many people who have age between 25-45 and 45-65 are affected by the Melanocytic Nevi skin cancer, and mostly female peoples are affected by this kind of disease.

4.8 Vascular Lesions – Localization

Table 4.13 Vascular Lesions -Localization based Agewise Report

Localization	Age	Gender		
		Male	Female	Total
Scalp	5-25	-	-	-
	25-45	1	-	1
	45-65	1	-	1
	65-85	-	-	-
Ear	5-25	-	-	-
	25-45	-	-	-
	45-65	-	-	-
	65-85	-	-	-
Face	5-25	2	-	2
	25-45	1	-	1
	45-65	3	3	3
	65-85	-	2	2
Back	5-25	3	5	8
	25-45	-	-	-
	45-65	2	5	7

	65-85	7	-	7
Trunk	5-25	4	-	4
	25-45	5	3	8
	45-65	12	4	16
	65-85	-	-	-
Chest	5-25	2	5	7
	25-45	-	-	-
	45-65	74	66	140
	65-85	-	1	1
Upper extremity	5-25	5	-	5
	25-45	9	8	17
	45-65	8	2	10
	65-85	-	-	-
Unknown	5-25	2	5	7
	25-45	-	-	-
	45-65	74	66	140
	65-85	32	-	32
Abdomen	5-25	2	5	7
	25-45	-	-	-
	45-65	9	8	17
	65-85	1	5	6
Lower extremity	5-25	7	6	13
	25-45	74	66	140
	45-65	-	-	-
	65-85	2	5	7
Genital	5-25	88	33	121
	25-45	-	-	-
	45-65	21	-	21
	65-85	9	8	17
Neck	5-25	4	3	7
	25-45	-	-	-
	45-65	1	6	7
	65-85	74	66	140
hand	5-25	-	-	-
	25-45	2	5	7
	45-65	88	33	121
	65-85	-	-	-
Foot	5-25	3	7	10
	25-45	9	8	17
	45-65	-	-	-
	65-85	74	66	140

Table 4.13 shows the age wise report of people who affected by the skin cancer based on localization attribute. Results Observed by the performance of the algorithms. From this analysis many people affected by the Vascular Lesions skin cancer disease on their chest, abdomen and foot. And mostly many people who have age between 5-25 and 45-65 are affected by the Basal cell skin cancer skin cancer, and mostly male peoples are affected by this kind of disease.

4.9 Dermatofibroma – Localization

Table 4.14 DF-Localization based Age wise Report

Localization	Age	Gender		
		Male	Female	Total
Scalp	5-25	-	-	-
	25-45	-	-	-
	45-65	-	-	-
	65-85	-	-	-
Ear	5-25	2	-	2
	25-45	-	-	-
	45-65	-	-	-
	65-85	-	-	-
Face	5-25	-	-	-
	25-45	-	-	-
	45-65	-	-	-
	65-85	-	-	-
Back	5-25	2	-	2
	25-45	-	-	-
	45-65	-	-	-
	65-85	-	-	-
Trunk	5-25	-	-	-
	25-45	-	-	-
	45-65	-	-	-
	65-85	-	-	-
Chest	5-25	4	3	7
	25-45	2	5	7
	45-65	-	-	-
	65-85	5	-	5
Upper extremity	5-25	-	8	8
	25-45	88	33	121
	45-65	2	5	7
	65-85	4	3	7
Unknown	5-25	-	-	-
	25-45	-	-	-
	45-65	-	-	-
	65-85	-	-	-
Abdomen	5-25	-	-	-
	25-45	-	-	-
	45-65	9	8	17

	65-85	2	5	7
Lower extremity	5-25	7	5	12
	25-45	6	3	9
	45-65	-	-	-
	65-85	-	-	-
Genital	5-25	-	-	-
	25-45	-	-	-
	45-65	-	-	-
	65-85	-	-	-
Neck	5-25	-	-	-
	25-45	-	-	-
	45-65	-	-	-
	65-85	-	-	-
Hand	5-25	2	5	7
	25-45	74	66	140
	45-65	21	10	31
	65-85	4	3	7
Foot	5-25	-	-	-
	25-45	-	-	-
	45-65	6	2	8
	65-85	74	66	140

Table 4.14 shows the age wise report of people who affected by the skin cancer based on localization attribute. Results Observed by the performance of the algorithms. From this analysis many people affected by the Dermatofibroma skin cancer disease on their upper extremity, lower extremity and hand. Most of people who have age between 25-45 and 45-65 are affected by the Dermatofibroma skin cancer skin cancer, and mostly female peoples are affected by this kind of disease.

4.10 Akiec – Localization

Table 4.15 Akiec -Localization based Age wise Report

Localization	Age	Gender		
		Male	Female	Total
Scalp	5-25	-	-	-
	25-45	-	-	-
	45-65	4	-	4
	65-85	10	-	10
Ear	5-25	-	-	-
	25-45	-	-	-
	45-65	2	-	2
	65-85	1	-	1
Face	5-25	-	-	-
	25-45	9	2	11
	45-65	32	32	64
	65-85	24	14	38
Back	5-25	-	-	-
	25-45	-	-	-

	45-65	12	-	12
	65-85	9	5	14
Trunk	5-25	-	-	-
	25-45	1	-	1
	45-65	-	-	-
	65-85	-	-	-
Chest	5-25	4	3	7
	25-45	9	8	17
	45-65	2	5	7
	65-85	8	5	13
Upper extremity	5-25	74	66	140
	25-45	-	-	-
	45-65	88	33	121
	65-85	-	-	-
Unknown	5-25	4	2	6
	25-45	8	-	8
	45-65	9	8	17
	65-85	4	3	7
Abdomen	5-25	-	-	-
	25-45	2	5	7
	45-65	74	66	140
	65-85	6	2	8
Lower extremity	5-25	88	33	121
	25-45	112	-	112
	45-65	9	8	17
	65-85	23	-	23
Genital	5-25	74	66	140
	25-45	20	21	41
	45-65	2	5	7
	65-85	88	33	121
Neck	5-25	9	4	13
	25-45	-	-	-
	45-65	5	-	5
	65-85	9	8	17
Hand	5-25	7	2	9
	25-45	-	-	-
	45-65	-	3	-
	65-85	74	66	140
Foot	5-25	-	-	-
	25-45	88	33	121
	45-65	4	9	13
	65-85	-	-	-

Table 4.15 shows the age wise report of people who affected by the skin cancer based on localization attribute. Results Observed by the performance of the algorithms. From this analysis many people affected by the Akiec skin cancer disease on their lower extremity, abdomen and hand. Most of people who have age between 25-45 and 65-85 are affected by the Akiec skin cancer skin cancer, and mostly female peoples are affected by this kind of disease**.**

The performances of these algorithms are measured by using classification accuracy, error rate and time. Neural Network Algorithm shows better accuracy than other algorithms such as Random Forest, Recursive Partitioning, SVM-Linear, SVM-Kernel. Finally, in this age wise report, the age is categorized into four groups. They are 5-25, 25-45, 45-65, 65-85. In the 5-25 age wise report observed totally 601 people are affected in which 339 female and 262 male. In the 25-45 age wise report observed totally 525 people are affected in which 313 female and 212 male. In the 45-65 age wise report observed totally 3733 people are affected in which 1652 female and 1652 male. In the 65-85 age wise report observed totally 2072 people are affected in which 645 female and 1427 male. From Localization based age wise report, totally 601people are affected by the Benign Keratosis-Like Lesions, here most of the people affected on their face and hand. In the Basal cell carcinoma disease 525 people are affected, most of the people affected on their chest, upper extremity and foot. In the Basal cell carcinoma disease 525 people are affected, most of the people affected on their chest, upper extremity and foot. In the melanoma disease 1625 people are affected, most of the people affected on their chest, lower extremity, foot and back. In the Melanocytic Nevi disease 3635 people are affected, most of the people affected on their upper extremity, abdomen, foot, trunk and back. In the Vascular Lesions disease 365 people are affected, most of the people affected on their lower extremity, foot, chest and neck. In the Dermatofibroma disease 275 people are affected, most of the people affected on their upper extremity and hand. In the Akiec disease 875 people are affected, most of the people affected on their lower extremity and genital. Finally that the highest count of female who have an age between 25-65 are affected by the type of skin cancer such as Melanocytic Nevi.

5. CONCLUSION AND FUTURE WORK

The main aim of the research work is to classify the skin data set using data mining classification algorithms such as Random Forest, Recursive Partitioning, SVM-Linear, SVM-Kernel and Neural Networks. The performance of the algorithms is measured by using correctly classified instances and incorrectly classified instances. The skin cancer dataset contains 2017-2019 two years information which includes 10015 instances and 8 attributes such as lesion_id, image_id, dx, dx_type, age, gender, localization and class. Here Neural Network Algorithm shows better accuracy than other algorithms such as Random Forest, Recursive Partitioning, SVM- Linear and SVM-Kernel. The algorithm which has the highest accuracy is chosen as the best algorithm. By considering different parameters of accuracy, it is found out that the Neural Network classification algorithm is the best algorithm with a maximum accuracy of 98.3 than other classification algorithms.Finally, observed the most number of female who have an age between 25-65 are affected by the type of skin cancer such as Melanocytic Nevi . In future, machine learning and optimization algorithms will be considered for better accuracy.

REFERENCE

[1] Divya Jain, Vijendra Singh, " A review Feature selection and classification systems for chronic disease prediction", Egyptian Informatics Journal, April 2018.

[2] Vikas Chaurasia, Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques ", Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, Mar 2017.

[3] Meherwar Fatima, Maruf Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic", Journal of Intelligent Learning Systems and Applications, September 2017.

[4] V. Krishnaiah, G. Narsimha, N. Subhash Chandra," A review of Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach", International Journal of Computer Applications, February 2016.

[5] R.Thanigaivel, Dr. K.Ramesh Kumar, "Review on Heart Disease Prediction System using Data Mining Techniques ", Journal of Intelligent Learning Systems and Applications, April 2015

[6] Dr. S. Vijayarani,Mr.S.Dhayanand, "DATA MINING CLASSIFICATION ALGORITHMS FOR KIDNEY DISEASE PREDICTION", International Journal on Cybernetics & Informatics (IJCI), August 2015.

[7] Dr. S. Vijayarani,Mr.S.Dhayanand," Liver Disease Prediction using SVM and Naïve Bayes Algorithms", International Journal of Science, Engineering and Technology Research (IJSETR),April 2015

[8] T. Sathya Devi, Dr. K. Meenakshi Sundaram, "A Comparitive Analysis of Meta And Tree Classification Algorithms Using Weka".International Journal of Engineering and Technology Volume: 03 Issue: 11/NOV-2016

[9] Nanak Chand, Preeti Mishra, "A Comparative Analysis of SVM and its Stacking with other Classification Algorithm for Intrusion Detection". Conference paper April 2016

[10] Songul Cinaroglu "Comparison of performance of decision trees and random forest An Application on OECD Countries Health Expenditures". International Journal of computer Application, Volume 138, March 2016.

[11] Bharathi, M.Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering Vol. 1 NO.4301-305.

[12] Sudhamathy.G, Thilagu.M, "Comparative Analysis of R Package Classifiers using Breast Cancer Data set", International Journal of Engineering and Technology, Volume 8 Issue 5, Oct-Nov-2016.

[13] Prachi Damodhar Shahare, Ram Nivas Giri, "Comparitive Analysis of Artificial Neural Network and Support Vector Machine for Classification of Breast Cancer Detection", International Research Journal of Engineering and Technology, Volume 2, Issue 9,Dec 2015.

[14] Parul Sinha, Poonam Sinha, "Comparitive Study of Chronic Kidney Disease Prediction using KNN and SVM", International Journal of Research Engineering and Technology volume 4 Issue 12 Dec 2015.

[15] https://en.wikipedia.org/wiki/Support_vector_machine

[16] Lakshmi Devasena C, "Comparitive Analysis of Random Forest, REP Tree and J48 Classifiers for Credit Risk Prediction", International Journal of Computer Application, 2014.
[17] https://www.datasciencecentral.com/profiles/blogs/random-forests-algorithm
[18] http://www.exforsys.com/tutorials/data-mining/data-mining-applications.html
[19] https://www.google.com/search?rpart+algorithm+in+rstudio+on+wikipedia+&oq=rpart+algorithm+in+rstudio+on+wikipedia
[20] http://www.dictionary.com/browse/data-mining
[21] https://en.wikipedia.org/wiki/Recursive_partitioning
[22] https://eight2late.wordpress.com/2016/02/16/a-gentle-introduction-to-decision-trees-using-r/
[23] https://simple.wikipedia.org/wiki/Data_mining
[24] https://www.listendata.com/2014/11/random-forest-with-r.html
[25] https://www.guru99.com/r-random-forest-tutorial.html
[26] https://www.google.com/search?rlz=1C1CHBF/ neural+network+tutorial+in+rstudio
[27] http://dataaspirant.com/2017/01/13/support-vector-machine-algorithm/
[28] P.Ramachandran, " Early Detection and Prevention of Cancer using Data Mining Techniques", International Journal of Computer Applications (0975 – 8887) Volume 97– No.13, July 2014.
[29] K.Arutchelvan, "Cancer Prediction System Using Data mining Techniques", International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 08 | Nov-2015 .
[30] Rahul Krishna Kota, "Diagnosing Common Skin Diseases using Soft Computing Techniques", International Journal of Bio-Science and Bio-Technology Vol.7, No.6 (2015).