CLASSIFICATION OF SKIN CANCER DISEASE USING DATA MINING TECHNIQUES

BY

SHADMAN SHOWMIK

152-15-5819

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

MD ZAHID HASAN

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2019

APPROVAL

This Project titled **"Classification of skin cancer disease using data mining techniques,"** submitted by Shadman Showmik, ID No: 152-15-5819 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 03 May 2019.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Ismail Jabiullah — **Internal Examiner**
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Dewan Md. Farid — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
United International University

© Daffodil International University ii

DECLARATION

I hereby declare, this project has been done by me under the supervision of **Md. Zahid Hasan, Assistant Professor, Department of CSE** Daffodil International University. I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Md. Zahid Hasan

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Submitted by:

Shadman Showmik

Shadman Showmik

ID: 152-15-5819

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

ACKNOWLEDGEMENT

I have given my efforts to this thesis. However, it would not have been possible without the kind support and help of many individuals. I would like to express my deepest appreciation to all those who provided me the possibility to complete this report.

At first, I express my heartiest thanks and gratefulness to almighty Allah for His divine blessings which allowed me to complete this thesis successfully.

A special gratitude I give to my supervisor, **Md. Zahid Hasan**, Assistant Professor of CSE department, whose contribution in stimulating suggestions and encouragement helped me to coordinate my thesis especially in writing this report. His endless patience, scholarly guidance, constant and energetic supervision, constructive criticism, valuable advice have made it possible to complete this thesis.

Furthermore, I would also like to acknowledge with much appreciation the crucial role of my department head, **Professor Dr. Syed Akhter Hossain**, who provided me with his precious time and kind help to finish this thesis. I also give my deepest thanks to all the faculty members and staff of CSE department of Daffodil International University.

ABSTRACT

In the previous decade fast growing of digital facts and worldwide availability of it over current internet has perceived a huge increase in machine learning exploration. In ratio to that, the health data has similarly seen a huge range of development. By the obtainability of organized clinical data, it has involved researchers to learn on the computerization of medical disease discovery through machine learning and data mining. Melanoma is a fatal skin malignance that breakdowns in the skin's tincture cells on the membrane shallow. Melanoma origins 75% of the skin cancer-associated deaths. This disease be able to identify by a dermatology expert over the clarification of the dermoscopy imageries in keeping with ABCD law. So, our investigation goals to study the automated discovery of skin cancer ailment through medical data by numerous machine learning classifier. This exploration mainly emphases on Neural Nets, Deep learning, Naïve Bayes, Random Forest classifier and decision tree in the determination of categorizing the intended dataset in three groups as normal, abnormal and melanoma to develop a decision support system that would create the assessment easier for a doctor. Generally our attempt has been to attain a supportable and realistic model to distinguish the skin cancer disease through comprehensive scientific accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES **PAGE**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Till now Cancer is an unavoidable issue for scientific community reason for no current medicines could tackle the issue related to this horrible disease. Research is in well advancement since 50 years yet it abandoned to give an exact answer for fight against it. The improvement of innovation in science day night endeavors to grow new techniques for treatment. Malignancy is an ailment that happens through multiplying of the body cells in an uncontrolled way and possessing the peripheral tissues.

Skin is that, which serves to shields the body from disease, warmth, damage, and any kind of harm which brought about by ultraviolet (UV) radiation, (for example, from the sun or sunlamps), germs, compound. Skin stores water and fat. It encourages body to control warmth and water. Likewise, encourages skin to make vitamin D. So protection of skin from ailment is the critical and unpredictable activity in treatment. At some point skin is influenced by various types of diseases where skin cancer is generally one. Skin diseases are named for the kind of cells that end up threatening (malignant growth).

Despite the fact that the skin malignant growth disease happens less continuous than numerous other cancer types, it is exceptionally vital as a result of its high mortality. The skin disease has distinctive of types, for example, Malignant Melanoma, Squamous cell carcinoma, and Basal cell carcinoma. The characteristic advancement of melanoma happens in two phases aside from the nodular kind. The horizontal improvement organize that advances along the epidermal surface, is characterized as "single cancer melanoma" which has a basic significance for the early analysis.

Dermoscopy is a non-intrusive symptomatic strategy that enables us to inspect in more detail the morphological structure of the pigmented skin injuries. The melanoma finding is performed by interpreting the pictures obtained with the dermoscopy method. The dermoscopy method permits through the harsh light the detailed imagining perception of the morphological structures and examples. Dermatologists ordinarily execute the determination of melanoma through these pictures by ABCD

(Asymmetrical Shape, Border, Color and Diameter) rule. ABCD is a very subjective evaluation that is reliant on the experience and learning of the related specialists [1].

In this way, the scope of this exploration is to construct a model utilizing information mining procedures to predict if a patient does to be sure have melanoma or not by evaluating and examining indications and different health parameters; Using Data Mining strategies to classify those information and think about the consequences through various methods.

1.2 Motivation:

Melanoma occurrence is being accounted to rise more quickly than other types of cancer. Melanoma is in charge of 4% just of all skin cancers, while it is in charge of 75% aggregate of skin cancer deaths. Melanoma, which is supposed to be stimulated by ultraviolet rays, is more usually happening in area where exposure to daylight is relative higher. In Europe are distinguishing 62.000 new cases every year. As indicated by the American Cancer Society's report 2016 was for the year 2016 predicted that 76.380 cases will be determined to have melanoma in the United States and 10.130 individuals will die from it [2].

Melanoma is a brutal skin disease that breaks out in the skin's pigment cells on the skin surface. Melanoma reasons 75% of the skin disease allied deaths. This disease can be analyzed by a dermatology expert through the clarification of the dermoscopy pictures as per ABCD rule. Regardless of whether dermatology specialists utilize dermatological pictures for determination, the rate of the right analysis of specialists is assessed to be 75-84% [3].

In Bangladesh, vast portion of general public living underneath the poverty line does not have adequate access to the required medical attention. The public division which is totally kept running by the government subsidize does not have the money related capacity or the best possible restorative assets to consolidate this expansive amount of fiscally unprivileged people to the therapeutic sector. Thus it is a very significant health concern in the skin disease. The circumstance is worsening as the high expense of diagnosing test also accountable for people turning away to get tested for skin

disease. Subsequently a major part of individuals bearing the disease are absolutely careless in regards to their own skin issue.

As the initial move towards treatment in any medical condition is to getting analyzed first. With the progress of healing innovation and ability of storing medical information in computerized shape has revived the possibility of medicinal computerization and information revolution has everything except made the possibility of artificially programmed doctor something other than an ambitious dream. A computerized virtual framework to classify skin malignant growth disease, is yet not so much convincing to the vast main stream of specialists and medical individual. But with more information, productivity and more exactness, a future of mechanized artificial medical aide can turn into a reality.

As the universe of innovation is moving towards electronic mechanization and AI automation raw data is being created continuously. This bounty of information abandons us with the chance to analyze those information with new data mining and machine learning methods.

So to the extent as the medicinal information is concerned increasingly more so computerized innovation and preservation of patient's information is becoming gradually common. Investigating those information with conventional measurements may give us the why's yet utilizing machine learning and data mining can show the potential solution of those issues. Doing classification by machine learning, deep learning and data mining is being trained progressively. These methods can also uncover unknown patterns [4].

From Bangladesh's perspective the skin cancer circumstance can be a field of study by means of computerized classification as the disease is genuinely not hard to classify applying appropriate information. The public sector works with vast number of patient at the same time. So an automated method can give a pre-indicative suggestion by classification of the test outcomes which a specialist doctor can affirm after checking. This testing stage can be executed for a specific timeframe amid which further information will be gathered. Further preparing of the model via continuous and genuine clinical information will be a dive forward to realizing how much we would be able to depend on an automated system. What's more, with the

Increasing exactness and consistency there is a genuinely decent possibility of computerizing skin cancer diagnosis and a many more ailments with appropriate implementation.

1.3 Rationale For the study

Here are a certain reasons and disputations for the investigation on skin cancer classifications.

- The prospective of an automated framework for classification of disease has continually attracted scientists. In spite of that medical acceptance for this has not acquired too much consideration, it certainly has potential to get executed.

- Skin Cancer disease can be tricky in nature as the side effects comes frequently at a late stage. The system can advise the patients which test to take and once the test is done the patient can check their own skin disease risk dimension themselves.

- An automated method can continually check the risk level of the patients in the public hospitals later which can be confirmed by the doctors.

- The overwhelming number of individuals that look for medicinal services in the public hospitals in Bangladesh has to bear endless suffering to go through the process. An automated system can definitely save those people a great deal of money and time. The hospital be able to filter out lots of patient who does not require medical consideration by using the system, leaving just the individuals who are really at danger of the disease. This will spare time, assets and money for the public hospitals.

- With regard to Bangladesh, Skin Cancer is harmful and pricey disease whose chance is expanding at alarming rate due to unconsciousness. And what makes the things worse is that lot of them rely on public treatment where it has deal with this immense number of patient resulting in long procedure of health services.

1.4 Research Question

- ❖ Can the skin cancer disease be classified with Convincing level of medical accuracy?
- ❖ Applying different classifier algorithm to the dataset.
- ❖ Comparing and enlisting the accuracy rate of every classifiers.
- ❖ As the dataset is quite general we will work of the dataset with different subset.

1.5 Expected Outcome

- Comparing different Machine Learning classification methods.
- By analyzing the dataset skin cancer status of patient should be classified with comprehensive accuracy.
- To provide the how exactly these data mining algorithms can predict the earlier warning to the users.
- This system estimates the risk of the melanoma and non-melanoma skin cancers.
- Finding out the correlation of different attributes in the dataset in developing skin cancer.

1.6 Report Layout

- In the first chapter of the project report we have discussed about the overview and motivation of our project. We have also discussed of our objective and expected result.
- The second chapter we have widely discussed about our background study on skin cancer disease and literature review. We have also enlisted number of studies on this field. The research methodology have described in the third chapter that we have used. In this research we have also briefly talked about the classifier algorithm.
- Fourth chapter includes experimental result's detailed description and related studies of the classifier algorithm accuracies. In the Fifth chapter we discussed about the summary, future scope of the study and further study territories in the similar field.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction:

In the following some parts we will discuss about skin cancer disease and it's Definition. We will explore the cause and risk of the skin cancer disease. We will also discuss about the literature survey in related field.

2.2 Skin Cancer Disease

2.2.1 Definition

Skin cancers are diseases that emerge from the skin. They are because of the growth of abnormal cells that can attack or spread to different parts of the body. Here exist three fundamental kinds of skin cancer: basal-cell skin cancer, squamous-cell skin cancer and melanoma as I previously said that. The initial two and together with numerous less frequent skin cancers, are recognized as skin cancer of non-melanoma. Basal-cell malignant growth matures slowly and can damage the skin around it though it is perhaps not going to extent too far off parts or else consequence in decease. It frequently seems as per an effortless elevated territory of skin that might be glossy by minor veins running over it or may present as an upraised area with an ulcer. Squamous-cell skin disease is more probable to spread. It generally presents as a hard inflammation with a flaky top but can also form an ulcer. Melanomas are the most destructive. Signs contain a mole that has changed in irregular edges, size, color, shape, is irritated or bleeds [5].

2.3 Cause and Risk of Skin Cancer

When one affected by cancer, it's normal to wonder what might have produced the disease that means what is the hazard or risk factor. Risk factor implies those reasons which raise the threat-chance of receiving an illness. The primary risk issue for skin cancer is disclosure to sunlight (Ultraviolet rays). UV radiation

causes around 65% to 90% Skin cancer. There are also some other risk elements of skin cancer. The Skin Cancer diagnosis is an important and tedious work. The identifications of Skin Cancer from some vital risk factors is a multi-layered issue.

Kids have a bigger number of chances and time than grown persons to be exposed to daylight and in this way more chances to build their danger of creating skin malignant growth. At least 25% of an individual's lifetime UV disclosure happens in the period of childhood. Most guardians recalled hearing about the significance of keeping their offspring from the sun, but kids are yet playing in the sun devoid of sunscreen or protective clothing. Many individuals nowadays are not considering skin cancer important subsequently knowing its consequences [6].

Skin malignance is increasingly regular when sun is robust. Daylight may be replicated by water, snowfall, ice and silt. Sun's beam might develop by mists, wind shields and light apparel. Thus we inspire people toward confine their exposure to daylight.

Another hazard issue of Skin cancer along with other maladies is Obesity. A person who has more weight than height that means more BMI is responsible for occurring different kinds of disease.

An individual who was influenced by any sort of malignant growth has an expanded danger of building up another skin disease of any kind. An individual who has at least two close relatives (mother, father, sister, and sibling) who are in charge of creating Skin disease has a hazard factor of creating skin malignancy for his own. Rarely, individuals of a family will have a hereditary disorder that creates the skin increasingly sensitive towards sun and expands the danger of skin cancer. Consuming fair skin that tingles in the sun simply, blue or grim eyes, red or light hair, or numerous spots expands the danger of skin cancer.

In Bangladeshi population skin cancer evaluation which contains- age, sex, inherited situation, outside deeds, working hours in industries, shade of body/skin, food habit, severe experienced in youth, past health checkup, use of anti-oversensitive medications, smoking, fatness, genetic threat, atmosphere, too much alcohol, radiation treatment and usage chemical substance for body without deduction its quality.

2.4 Related works and Comparative studies

The unstructured system of medicinal data and its environment of existence geographically diverse it hard to automate the classification of the ailment that much complex. Generally those information has geographical, national and social prejudice. And frequently clinical indications for a disease differs over various regions. So the In spite of those difficulties there have been different studies in the field which attempted to characterize skin cancer with a scope of data mining strategies. Huge numbers of those has been breaking new grounds and has brought new plans to group and classify. A portion of those are talked below. The melanoma identification and diagnosis can be enhanced by the ABCD instruction based and computer aided systems. These methods usually comprise of the separate parts for the image subdivision and segmentation, feature mining, extraction and classification respectively. Studies directed in this field are as per the following:

Kamasak et al. dermoscopic images classified after dividing the dermoscopic images through mining the Fourier identifiers of lesion. 83.33% accuracy they acquired on diagnosing of the melanoma [7].

Romero et al. in his paper suggested a resolution for supporting dermatologists through the analysis of skin lesions. More explicitly, they planned and executed a two-class classifier that takes skin sores images marked as an input as malignant or benign. Their proposed methodology got the accuracy of 81.33% [8].

Fathima et al. recommended technique distinguishes among melanomas, nevi, Basal Cell Carcinoma, and Seborrheic Keratosis and estimated 4 texture structures. The system accomplished skin lesions classification with SVM and KNN classifiers algorithm and computes the accuracy of these two classifiers. The SVM classifier achieved the accuracy of 75.95% and 68.30% accuracy for KNN classifiers [9].

Moataz et al. trained on an artificial neural network genetic algorithm technique for early recognition of the skin cancer and acquired a sensitivity of 91.67% and a specificity of 91.43% [10].

2.5 Research Summary

So as the past literature review and research study demonstrates there has been good number of concentrates in this field. The examinations and studies has been genuinely positive in their own particular manner. This kind of mechanized and computerized classification problem-issue has been investigated on numerous other disease.

From studying different algorithm to making re-optimization to the existing algorithm to find better results, researchers has gone through many different ways. The noticeable factor is that although the accuracy has been quite good, yet we have not seen any real implementation of this processes. Probably the idea of consulting a computerized diagnosis system for a disease isn't as convincing as consulting a doctor for the public. But with more accuracy and some experimental periods, a fully automated diagnosis probably would be as normal as consulting a doctor.

2.6 The scope of this problem

The scope of this problem is to classify our dataset using different machine learning algorithms which includes training and testing the model. We will try to explore the correlation between the dataset attributes to find out there dependency on each other in the development of skin cancer disease.

In Bangladesh an automated diagnosis system would reduce the lengthy process in health care. With an improved symptoms analyzing algorithm, the system can suggest diagnostic test to the users hence reducing time and cost in big hospitals.

2.7 Challenges

The primary challenge for this thesis is to collect data on skin cancer disease in Bangladesh. The dataset we used is detailed and was well pre-processed. In contrast, finding this type of dataset is quite difficult in Bangladesh. Same patient's data are not kept in a structured way. So test results for the same person is hard to find collectively. Besides the tests are often done discreetly. And all the test for required data are usually not done. So the data is in incomplete form. So with this kind of data results in insufficient or biased training which will result in lower accuracy.

The type of data is also important for the training of the model. Quite often the patient affected in skin cancer disease comes at a very late stage and so the model is trained with the data in such a way in which its class value is classified start from common nevus then atypical nevus and then melanoma.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Data Mining is a technique where large volume pre-existing raw data in database is processed, or altered to needs and analyzed to reveal useful patterns and new relations among attributes for achieving various goals. Data mining is also called knowledge discovery in databases also known as KDD [11].

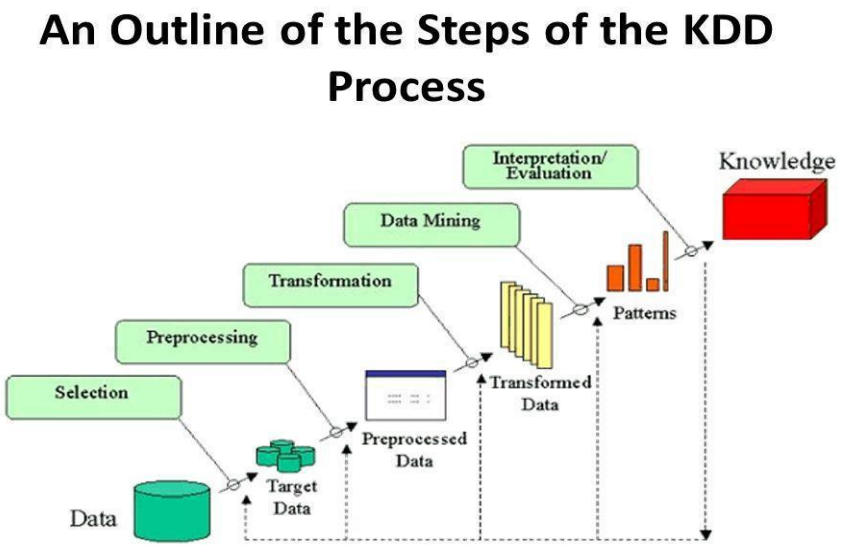

Figure 3.1: Steps associated with KDD

For a long time conventional data analysis techniques through statistical approach has been used. This approach has been very useful and no doubt it will be still be used in the foreseeable future As the storage capacity of modern computers increased, accumulating and preserving various transactional and other types of data became more convenient. Inevitably the size and diversity of the data grew larger and traditional data analysis techniques began to be less effective and inefficient for such large amount of data warehouses. So data mining and machine learning techniques gained popularity among large companies and researchers.

The latest explosion of Medical data through machine automation and use of computerized technology in diagnosis and treatment of disease has made the data mining and machine learning a “Gold Mine” for extracting new patterns and useful knowledge in medical advancement. Al though the acceptance of an automated classification of disease is still not popular and desirable among medical community, it is still a research area of enormous potential for data scientist and researchers around the globe.

So in our endeavor we will try to explore this concept of data mining to help automation of the classification of skin cancer disease.

3.2 About Dataset

In this research we used the data set which was developed at the Pedro Hispano Hospital of dermatology service by a group of researchers. They formed this PH2 dataset for melanoma diagnosis. The PH2 dataset contains 200 dermoscopy images at 768x560 resolution. Each depiction has RGB channels of 8 bit. The PH² database comprises medical explanation of all the imageries specifically medicinal segmentation of lesion, histological and clinical diagnosis and the evaluation of numerous dermoscopic criteria (Asymmetry; pigment network; regression areas; dots/globules; streaks; blue-whitish veil; colors) [12].

(a) Common nevus (b) Atypical nevus (c) Melanoma

Figure 3.2: Dataset Images

This data set is consist of three types of class value. First one Common nevus which refers to non-melanoma that means no skin cancer risk. Secondly Atypical

nevus which refers to risk of cancer and third one melanoma which indicates the cancer. The dataset has 80 instances of the class value for Common nevus and another 80 instances class value for atypical nevus and other 40 has been classified as melanoma.

3.3 Data Description and Preprocessing

In table 3.1 we have listed the attributes in the data set.

Table 3.1: - Attributes in the used dataset

Asymmetry	0 - Fully Simmetry
	1 – Asymmetry in One Axis
	2 - Fully Asimmetry
Pigment Network	AT - Atypical
	T - Typical
Dots/Globules	A - Absent
	AT - Atypical
	T - Typical
Streaks	A - Absent
	P - Present
Regression Areas	A - Absent
	P - Present
Blue Whitish Veil	A - Absent
	P - Present

Colors	1 - White
	2 - Red
	3 - Light-Brown
	4 - Dark-Brown
	5 - Blue-Gray
	6 - Black

1) Asymmetry: The asymmetry is a vital characteristic used for identifying a melanocytic lesion. It is very important in the ABCD rule of dermoscopy for its weight factor. The lesion asymmetry in this dataset was assessed by the clinician following to the ABCD rule. Hence, a lesion's asymmetry is evaluated regards to its shape, coloring and structure dissemination simultaneously. Besides, there are three feasible name for this parameter: 0 indicates fully symmetric; 1 for asymmetric with regard to one axis; and 2 used for fully asymmetric regarding two axes [12].

2) Colors: During the determination of a melanocytic lesion in total, six colors are considered into account. The color set consist of white, black, red, dark-brown, light-brown and blue-gray. Every image of dataset was assessed by a dermatologist so as to identify the existence, along with the location. The area of each color was noted as a binary mask in an image and manually segmented by the dermatologist [12].

3) Pigment network: It is a grid-like system comprising of hypo pigmented holes and pigmented line. This structure has a pivotal job in the qualification among melanocytic and non-melanocytic sores. The pigment network structure was outwardly assessed by the dermatologist, in each picture of the database, and named typical or atypical [12].

4) Dots/Globules: dots/globules are sphere-shaped or oval, differently sized, brown, gray or black structures (globules are usually greater than dots). The presence and existence of these dermoscopic structures is especially helpful for the difference among melanocytic and non-melanocytic lesions. These structures were outwardly assessed by dermatologists and classified categorically of the PH2 database as per present or absent in each image. Whenever these are present in a particular lesion, these structures are furthermore categorized as regular or irregular regarding their allocation in the lesion [12].

5) Streaks: Streaks are projections finger-like of pigment network from the fringe of the lesion. Rather than both dots/globules and pigment network, the available of streaks in a skin lesion is a sign of malignancy. Hence, these structures are simply categorized in each image of the database as present or absent[12].

6) Regression areas: Regression regions are characterized as white, scar-like depigmentation frequently joined with pepper like areas. In the PH2 database, this factor is characterized in two primary gatherings (present or absent) concerning its essence in the skin lesion [12].

7) Blue-whitish veil: Blue-whitish veil can be described as an irregular, confluent, hazy blue pigmentation with white, and ground-glass haze and an overlying. Its existence is an indication of strong malignancy. This dermoscopic structure and configuration is marked as present or absent, for each image of database [12].

Asymmetr	pigment n	dots/globu	streaks	regression	blue-whiti	white	red	light-brow	dark-brow	blue-gray	black
0	T	A	A	A	A	n	n	n	X	n	n
0	T	A	A	A	A	n	n	X	n	n	n
0	T	T	A	A	A	n	n	X	X	n	n
0	T	A	A	A	A	n	n	X	n	n	n
0	T	A	A	A	A	n	n	X	X	n	n
0	T	T	A	A	A	n	n	X	n	n	n
2	T	A	A	A	A	n	X	X	n	n	n
0	T	T	A	A	A	n	n	n	X	n	X
0	T	T	A	A	A	n	n	X	X	n	n
0	T	T	A	A	A	n	n	n	X	X	n
0	T	T	A	A	A	n	n	X	n	n	n
0	T	T	A	A	A	n	n	X	n	n	n
0	T	T	A	A	A	n	n	X	X	n	n
0	T	A	A	A	A	n	n	X	n	n	n
0	T	T	A	A	A	n	n	X	X	n	n
0	T	T	A	A	A	n	X	X	X	n	n
0	T	T	A	A	A	n	n	X	X	n	n
0	T	T	A	A	A	n	n	X	X	n	n
0	T	T	A	A	A	n	n	X	X	n	n
0	T	T	A	A	A	n	X	X	X	n	n

Figure 3.3: Dataset after pre-processing

3.4 Classification Algorithms

In ML classification or grouping the aim and vision is analyzing the training dataset to target the level class. Find the actual boundaries for every target class. In general by using the preparation dataset to acquire well margin states which might be salvaged to fix every goal class. Whenever the limit is defined, the following mission remains to target class predict. And this process is called classification. Here we use some of the classification methods to predict the class level.

3.4.1: Naive Bayes

Naive Bayes is a high-bias, low-variance classifier, and can be able to form a decent ideal model even with a minor information dataset. It is easy to utilize and computationally modest. Run of the mill use cases include content classification, including spam identification, assumption examination, and recommender frameworks. The principal suspicion is given estimation of the mark in class and estimation of every characteristic can autonomous of estimation in some further characteristic. Experience demonstrates that the Naive Bayes classifier frequently functions admirably. The freedom presumption boundlessly disentangles the computations expected to fabricate the Naive Bayes likelihood display. To finish the likelihood demonstrate, it is important to make some supposition about the restrictive likelihood disseminations for the individual Attributes, given the class. This Operator utilizes Gaussian likelihood densities to display the Attribute information [13].

3.4.2: Decision Tree

A decision tree is a tree like collection of nodes planned to make a resolution on standards association to a class or an estimation of a statistical objective value. Every hub connects towards a share rule on behalf of single explicit feature. For arrangement the standard separates standards taking an area through numerous classes, for relapse it isolates them so as to decrease the mistake in a perfect way aimed at the selected factor foundation. Structure of innovative hubs is repeated till the finishing standards are encountered. An expectancy on behalf of class designation feature is determined relying on most of instances which attained this sheet amongst age, whereas an estimate intended for statistical esteem be there acquired be an average of the qualities in a sheet. Name feature should be ostensible for arrangement and statistical intended for relapse. Afterward the choice tree typical model be able to connect. Every Example pursues the parts of the diagram in contract in part rule till a sheet node is reached [13].

3.4.3: Support vector machine

This learning method can be used for both regression and classification and stretches a rapid control and excessive consequences for some producing tasks. SVM mechanisms through quadratic or direct and sometimes misfortune capacities. The typical SVM takings a lot of info and forecasts, for every assumed material, double possible classes comprises the info, formatting SVM a non-probabilistic combined straight classifier. Assumed a lot of making instances, every single set separately as consuming an abode by single of two modules, a SVM fixing control productions a classical model that turns out innovative simulations into single grouping or other. A SVM reveal is a depiction of the prototypes as emphases in space, charted with the area that the examples of the dissimilar classes are inaccessible by a distinctive slum that is as extensive as could be permissible [13].

3.4.4: Random forest

A random forest is an ensemble of a certain number of random trees, determined through the amount of trees factor. Trees are prepared on sub-sets of the instance contribution ports. Each hub in tree states to a portion law for unique obvious characteristic. Equitable sub-set of features, specified with the subsection part base is deliberated for the portion decree optimal. Standard separates principles in a perfect pathway for the selected factor base. Grouping standard remains isolating qualities having a place with various classes, while for relapse it isolates them so as to diminish the blunder made by the estimation. The structure of fresh centers is repeated till the finishing norms are encountered. After age, irregular backwoods model can be connected. Every irregular tree produces a forecast for every Example by following the parts of the tree in agreement to the part administers and assessing the leaf. Class expectations depend on most of Examples, while estimations are gotten through the normal of qualities achieving a leaf. Essential parameters to tune for this technique are the insignificant leaf size

and split proportion, which can be changed in the wake of handicapping surmise split proportion. Great default decisions for the negligible leaf measure are 2 for grouping and 5 for relapse issues [13].

3.4.5: Deep learning

Deep Learning is based on a multi-layer feed-forward ANN that is qualified by stochastic incline gradient lineage using back-transmission propagation. The system can contain countless layers comprising of neurons with tanh, rectifier and maxout enactment capacities. Propelled highlights, for example, versatile learning rate, rate tempering, force preparing, dropout and L1 or L2 regularization empower high prescient exactness. Each process hub prepares a duplicate of the worldwide model parameters on its neighborhood information with multi-stringing (no concurrently), and contributes occasionally to the worldwide model by means of model averaging over the system [13].

3.4.6: Neural Net

This operator learns a model by means of a feed-forward neural network qualified through a back transmission procedure which is called multi-layer perceptron. The imminent units explain the vital views concerning neural systems and multi-layer perceptron. ANN as a rule is a numerical typical model that is brightened by the group and useful parts in natural systems. A neural system includes of a unified gathering of false neurons, and it procedures information using a connection method to contract with calculation (focal connection rule for psychological wonders may be depicted by consistent systems with basic and regularly constant elements). Much of the time an ANN is a useful outline that deviations its construction reliant on external or inner info that changes over the structure amid the knowledge phase. Present structures are usually accustomed to prove composite networks between bases of information and produces or to notice plans in data [13].

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In the following chapters we will discuss about the results of the conducted experiment. We will explore and compare the different classifier accuracy and performance. We will the results in graph and also in tables Receiver Operating Characteristics bended demonstrated both affectability and particularity of the test. The examination of TPR (True Positive Rate) and FPR (False Positive Rate) is characterized as ROC bend. The TPR is the extent of positive tuples that are effectively named by the model though FPR is of negative tuples misclassified as positive.

.

4.2 The Commonly-Accepted Performance Evaluation Measures

This is the case we focus on in this study. Classification performance without focusing on a class is the most general way of comparing algorithms. It does not favor any particular application. The introduction of a new learning problem inevitably concentrates on its domain, but omits a detailed analysis. Thus, the most used empirical measure and accuracy does not distinguish between the numbers of correct labels of the different classes [14].

TP = true positives: number of examples predicted positive that are actually positive FP = false positives: number of examples predicted positive that are actually negative TN = true negatives: number of examples predicted negative that are actually negative FN = false negatives: number of examples predicted negative that are actually positive

Accuracy: It refers to the total number of records that are effectively characterized by the classifier. Exactness of a classifier is characterized as the level of test set tuples that are effectively grouped by the model [15].

$$Accuracy=\frac{(TP+TN)}{(TP+FP+FN+TN)} \times 100\%$$

Sensitivity: Refers to the genuine positive rate that implies the extent of positive tuples that were effectively recognized.

Sensitivity=$\frac{TP}{TP+FN} \times 100\%$

Specificity: Indicates the rate at which a test or demonstrative technique sets a right (i.e. negative) finding for a patient who isn't sick [15].

Specificity=$\frac{TN}{TN+FP} \times 100\%$

Balanced accuracy (BACC): The balanced accuracy, which can be characterized as the normal precision acquired on either class.

BACC = $\frac{\frac{TP}{P}+\frac{TN}{N}}{2} \times 100\%$

Precision: The portion of recovered cases that are important.

Precision = $\frac{TP}{TP+FP} \times 100\%$

F- measure: The F- measure also refers to F measures that combined both the measures Precision and Recall as the harmonic mean.

F-measure =$\frac{2*precision*sensitivity}{precision*specifity} \times 100\%$

ROC: Receiver Operating Characteristics bended demonstrated both affectability and particularity of the test. The examination of TPR (True Positive Rate) and FPR (False Positive Rate) is characterized as ROC bend. The TPR is the extent of positive tuples that are effectively named by the model though FPR is of negative tuples misclassified as positive [15].

I.e. TPR = TP (TP+FN) and FPR = FP (FP+TN)

4.3 Experimental Results

Following Section from 4.3.1 - 4.3.5 extensively discusses the results from our study. Here we see the weight of the all attributes. Weight is defined as the global importance of each of the attribute for the value of the target class which is independent of the modeling algorithm that we used in the study. Table 4.1 displays the weight of different attributes.

Table 4.1: The Weight of different attributes

Row No.	Attribute	Weight
1	pigment network = AT	1
2	blue-whitish veil = A	0.690
3	Asymmetry	0.626
4	regression areas = A	0.563
5	blue-gray = n	0.536
6	dots/globules = AT	0.436
7	black = n	0.398
8	white = n	0.359
9	streaks = A	0.330
10	dots/globules = A	0.124
11	dark-brown = X	0.114
12	light-brown = X	0.106
13	red = n	0

With according to the label attribute the weight attributes means that an attribute is considered to be more relevant if that attribute weight is higher. The table shows some interesting facts.

The pigment network has the highest weight which is clearly obvious because pigment network measurement is one of first clinical diagnosis test done for any skin disease patients. The noticeable fact from the study is that the blue-whitish veil and Asymmetry count has a very high weight of 0.690 and 0.626 which can be a very significant indication towards skin cancer detection and also might be

suggestive of clinical attention of skin cancer. The other weight of the attributes are listed in the table 4.1.

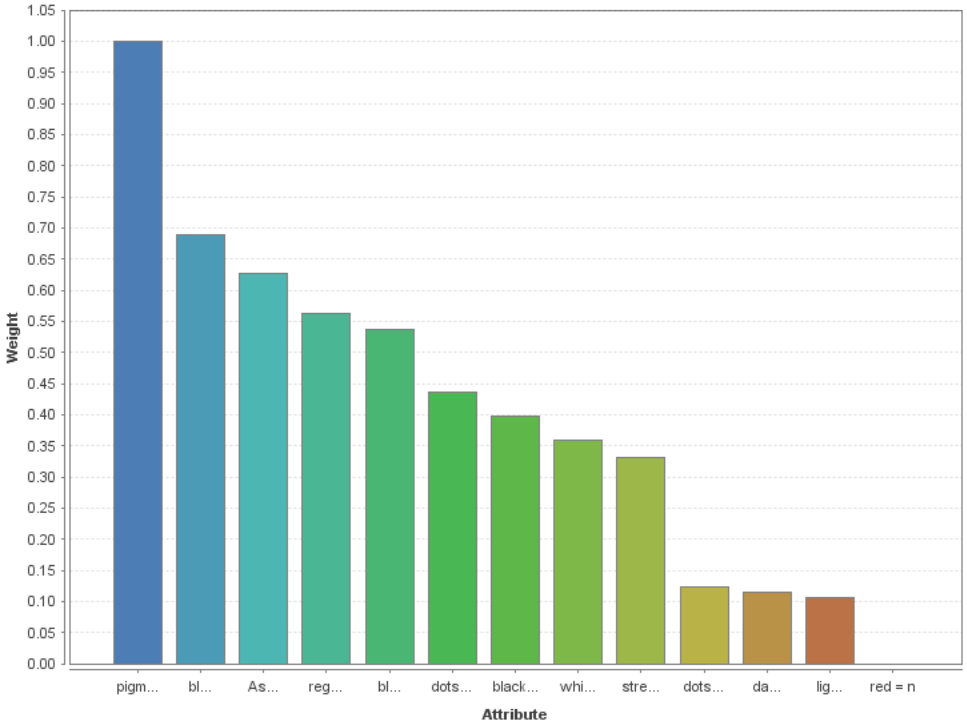

Figure 4.1: Weight of the attribute

4.3.1: Naive Bayes Confusion Matrix

Table 4.2: Confusion Matrix of Naive Bayes

Table View Plot View

accuracy: 91.67%

	true common nevus	true Atypical nevus	true Melanoma	class precision
pred. common nevus	23	0	0	100.00%
pred. Atypical nevus	1	22	2	88.00%
pred. Melanoma	0	2	10	83.33%
class recall	95.83%	91.67%	83.33%	

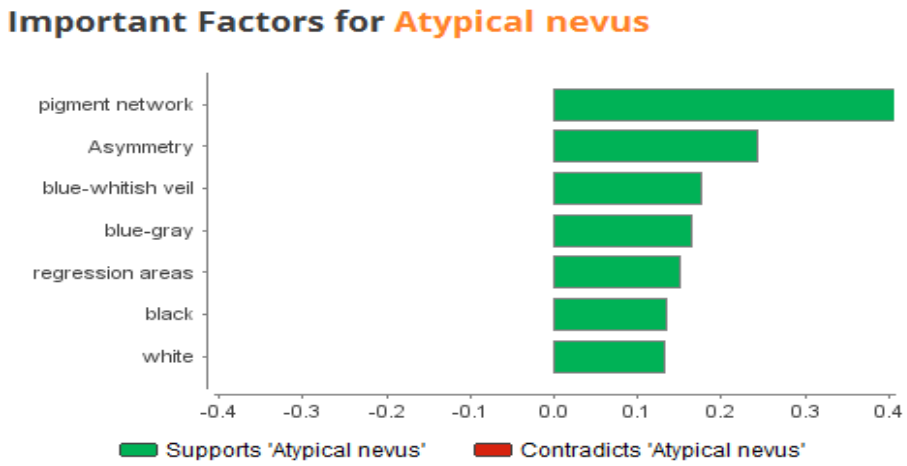

Figure 4.2: Important factors for Atypical Nevus in Naïve Bayes

4.3.2: Decision tree Confusion Matrix

Table 4.3: Confusion Matrix of decision tree

Table View Plot View

accuracy: 88.33%

	true common nevus	true Atypical nevus	true Melanoma	class precision
pred. common nevus	24	1	0	96.00%
pred. Atypical nevus	0	21	4	84.00%
pred. Melanoma	0	2	8	80.00%
class recall	100.00%	87.50%	66.67%	

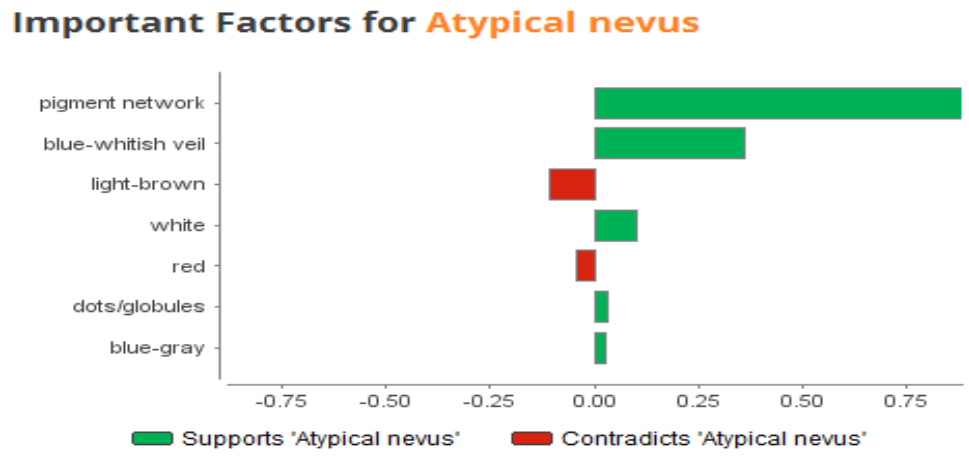

Figure 4.3: Important factors for Atypical Nevus in Decision Tree

4.3.3: Random forest Confusion Matrix

Table 4.4: Confusion Matrix of random forest

Table View Plot View

accuracy: 90.00%

	true common nevus	true Atypical nevus	true Melanoma	class precision
pred. common nevus	24	1	0	96.00%
pred. Atypical nevus	0	20	2	90.91%
pred. Melanoma	0	3	10	76.92%
class recall	100.00%	83.33%	83.33%	

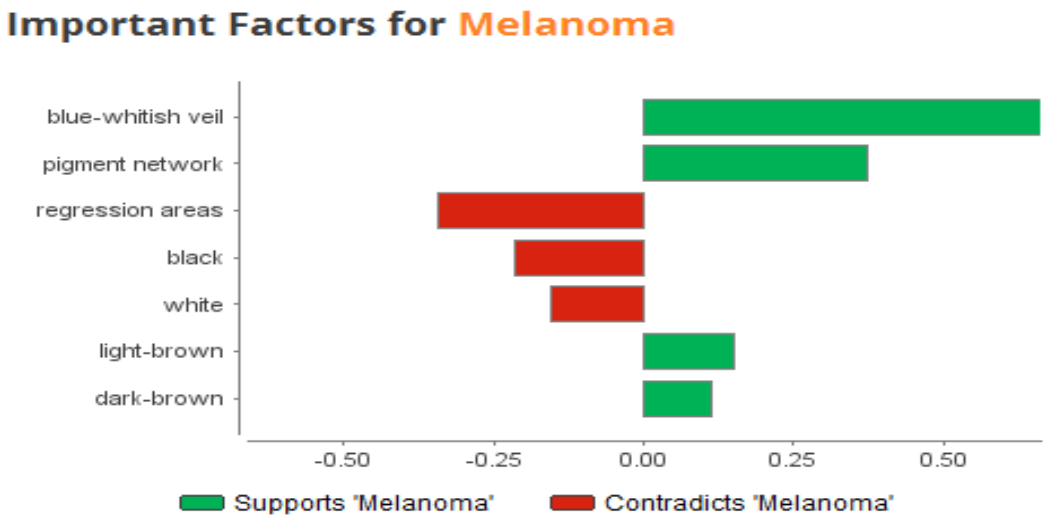

Figure 4.4: Important factors for Melanoma in Random Forest

4.3.4: Support Vector Machine Confusion Matrix

Table 4.5: Confusion Matrix of SVM

Table View Plot View

accuracy: 88.33%

	true common nevus	true Atypical nevus	true Melanoma	class precision
pred. common nevus	24	1	0	96.00%
pred. Atypical nevus	0	22	5	81.48%
pred. Melanoma	0	1	7	87.50%
class recall	100.00%	91.67%	58.33%	

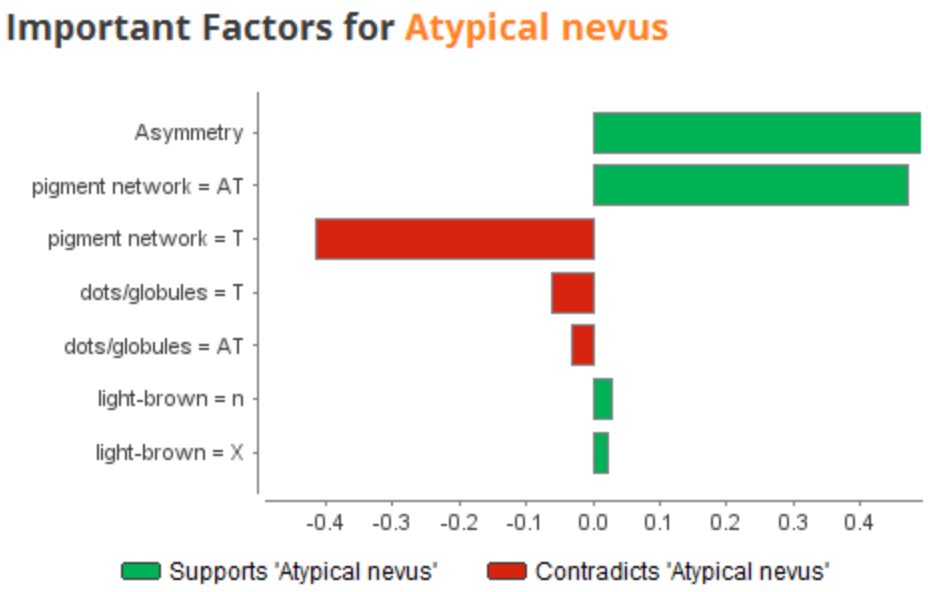

Figure 4.5: Important factors for Atypical Nevus in SVM

4.3.5: Neural Nets Confusion Matrix

Table 4.6: Confusion Matrix of neural Nets

Table View Plot View

accuracy: 93.33%

	true common nevus	true Atypical nevus	true Melanoma	class precision
pred. common nevus	24	1	0	96.00%
pred. Atypical nevus	0	21	1	95.45%
pred. Melanoma	0	2	11	84.62%
class recall	100.00%	87.50%	91.67%	

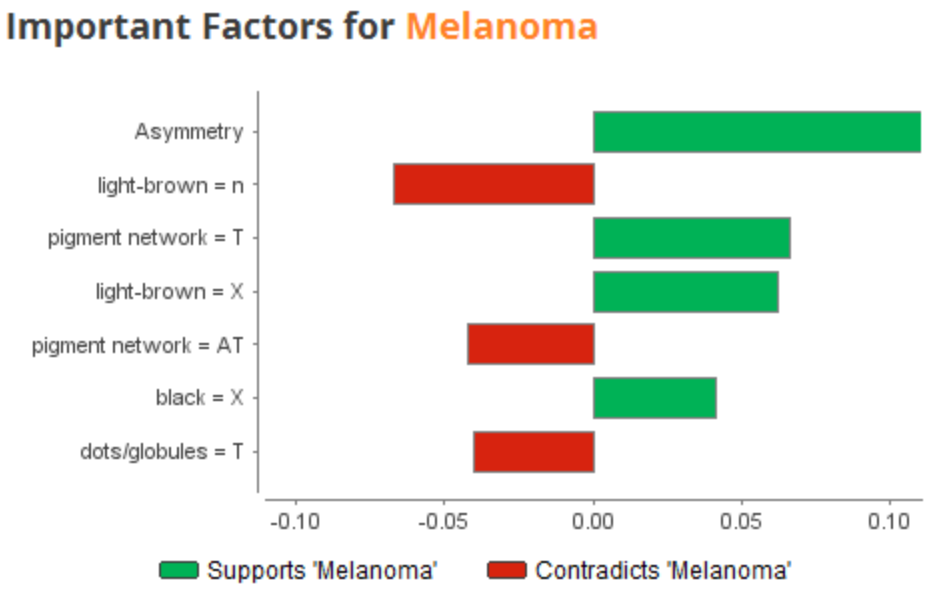

Figure 4.6: Important factors for Melanoma in Neural Nets

4.3.6: Deep Learning Confusion Matrix

Table 4.7: Confusion Matrix of Deep Learning

accuracy: 95.00%

	true common nevus	true Atypical nevus	true Melanoma	class precision
pred. common nevus	16	0	0	100.00%
pred. Atypical nevus	0	16	2	88.89%
pred. Melanoma	0	0	6	100.00%
class recall	100.00%	100.00%	75.00%	

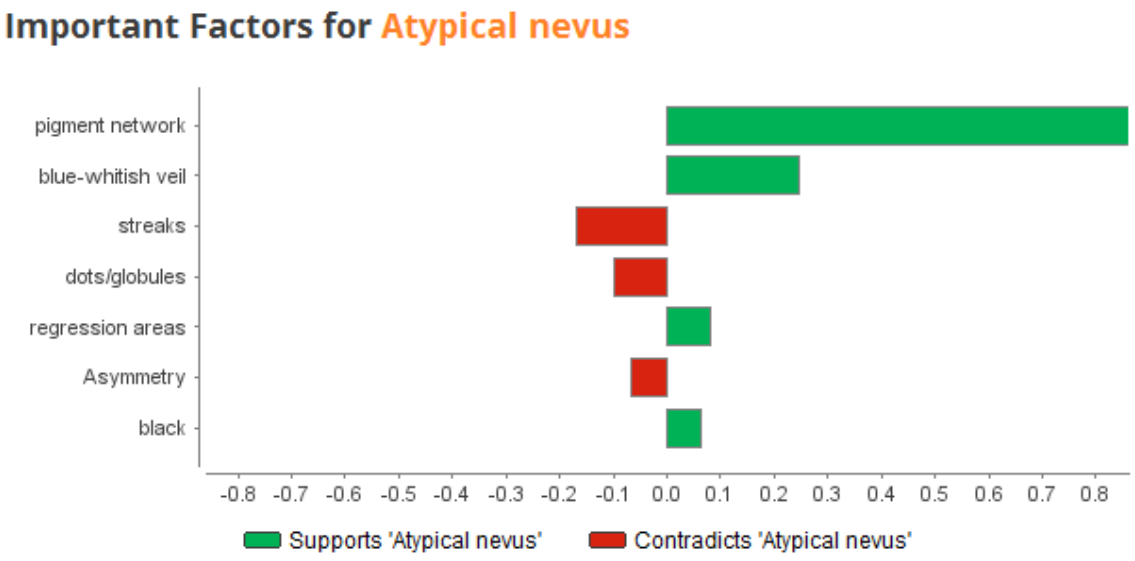

Figure 4.7: Important factors for Atypical nevus in Deep learning

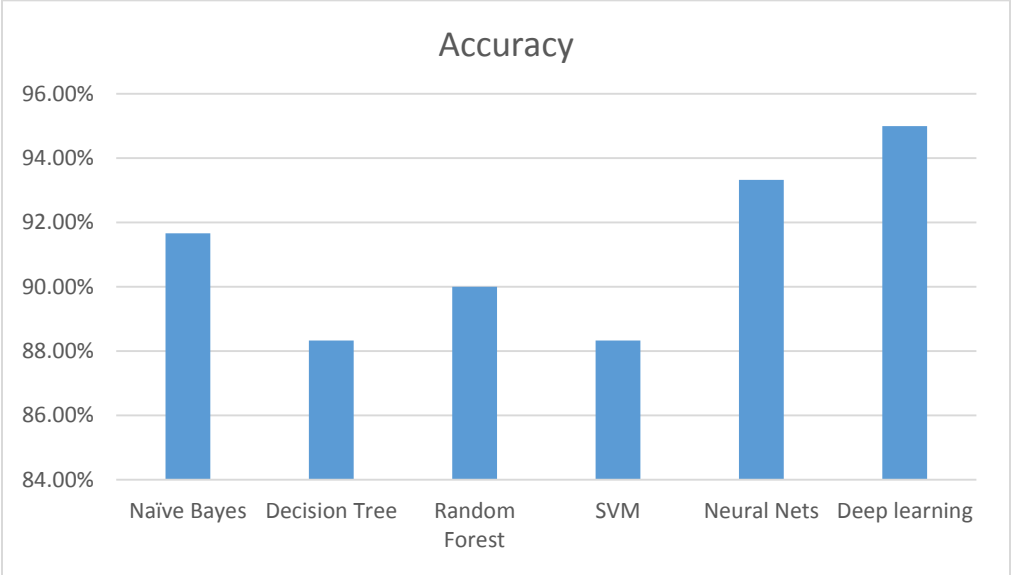

Figure 4.8: Accuracy of all classifiers

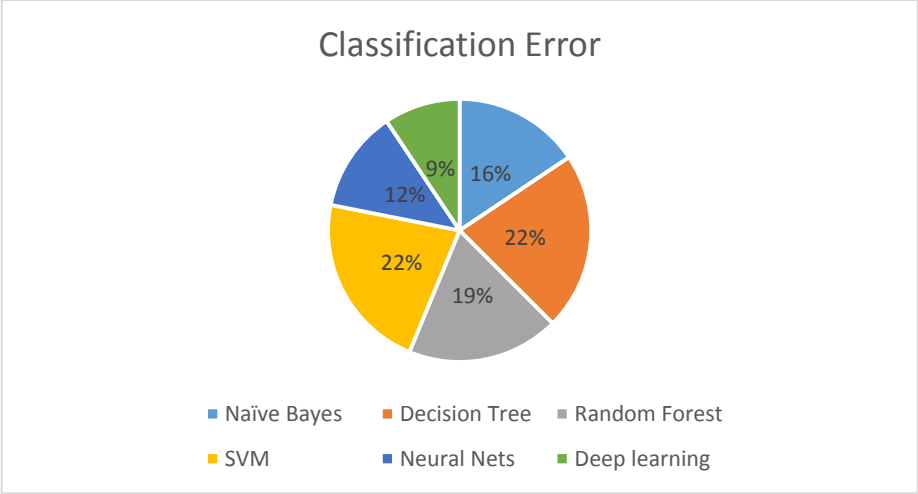

Figure 4.9: Classification Error

4.4 Potential Future Improvement

The study shows that an automated system can be implemented in the clinical diagnosis of Skin Cancer disease. With more training data and more efficient algorithm this can be a real life implementation for clinical diagnosis in Bangladesh.

Bangladesh medical system can accumulate more clinical data in an organized and structured way where each patient's medical data can be secured which can be later be used in many different studies for other diseases and also incorporate with the existing system for automated classification and detection of those diseases.

4.5 Discussion

So to conclude it all, we have used three different classifier among which the Deep learning classifier had the highest level of accuracy. Although the other classifier also gave very close and accurate result compared to Deep learning. So this Model can be used with convincing accuracy for a Clinical testing period to find out its feasibility and sustainability in practical use. More clinical data needs to be amassed with required data organization which should be used to train the model order for it to be used as a truly medically comprehensive platform for automated skin cancer disease detection.

CHAPTER 5

SUMMARY AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary

In brief the Deep Learning Classifier had the most accurate detection of the disease which is 95%. According to a study conducted by Jain et al., even though expert dermatologists use dermatology images for diagnosis, the rate of correct diagnosis of experts is estimated at 75-84%. In this study, which was performed with the different classification algorithms to classify the skin lesions, the normal skin lesions were by NN and DT classifiers more than 95% correctly classified according to the data acquired from PH2 data set.

All of the classifier algorithms used are revealed in terms of the classification outputs to be "better" than others in the normal type classification and to be "worse" than the others in the melanoma type classification. When the obtained data should be evaluated in terms of the output accuracy ratios and accuracy level of each class, will be observed that Deep Learning has more successful classified the PH2 data set than Naïve Bayes, Random Forest, SVM and DT. An accuracy of 95% achieved with the Deep learning classifier reveals that this classifier is a medical decision support system which could help dermatologists to diagnose the skin lesions.

5.2 Implication for further research

Additionally this study has distinguished with its higher accuracy, specifity and balanced accuracy ratio equated to other studies. This study maybe further progressed by using the different preliminary data processing techniques and hybrid classification algorithms. In addition, this study can be combined with the related image processing techniques also to be able to make autonomous decisions in several medical issues.

Further studies can be undertaken on various other diseases using similar techniques and more data on other clinical health problems should be accumulated in order for similar studies.

REFERENCES

[1] A. M. Y. Palomo, M. d. J. D. Pérez, O. R. P. Pérez, V. d. J. A. Yabor, and A. M. Fontaine, "Melanoma maligno cutáneo en pacientes de la provincia de Las Tunas," *Revista Electrónica Dr. Zoilo E. Marinello Vidaurreta*, vol. 40, 2015.

[2] Ahmed K., Jesmin T., Rahman M.Z. Early prevention and detection of skin cancer risk using data mining. Int. J. Comput. Appl. 2013; *62*(4).

[3] M. A. SedatÖzçelik, "Epidemiology of Melanoma," *TURKDERM*, vol. 41, pp. 1-5, 2007.

[4] 2016A. C. Society, "Cancer Facts & Figures," 2016.

[5] Stanley, R. Joe, Randy Hays Moss and Chetna Aggarwal. "A fuzzy based histogram analysis technique for skin lesion discrimination in dermatology clinical images", Computerized Medical Imaging and Graphics : the Official Journal of the Computerized Medical Imaging Society, Vol. 27, Number 5, pp. 387- 396, 2003.

[6] Ercal, Fikret, et al. "Neural network diagnosis of malignant melanoma from color images", IEEE Transactions on Biomedical Engineering, Vol.41,Number 9, pp. 837-845, 1994.

[7] E. Albay and M. Kamaşak, "Skin lesion classification using fourier descriptors of lesion borders," in Medical Technologies National Conference (TIPTEKNO), 2015, 2015, pp. 1-4.

[8] Adria Romero Lopez ; Xavier Giro-i-Nieto ; Jack Burdick ; OgeMarques,Skin lesion classification from dermoscopic images using deep learning techniques, IEEE Conference on Biomedical Engineering, 2017.

[9] Fathima Nizar, G.S Santhosh Kumar, “Classification of Various Skin Lesions using SVM and KNN Classifiers”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 8, August 2016.

[10] M. Aboras, H. Amasha, and I. Ibraheem, "Early detection of melanoma using multispectral imaging and artificial intelligence techniques," *American Journal of Biomedical and Life Sciences*, vol. 3, pp. 29-33, 2015.

[11] A. Baştürk, M. E. Yüksel, H. Badem and A. Çalışkan, "Deep neural network based diagnosis system for melanoma skin cancer," in *Signal Processing and Communications Applications Conference (SIU)*, 2017 25th, 2017, pp. 1-4.

[12] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "PH 2-A dermoscopic image database for research and benchmarking," in Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, 2013, pp. 5437-5440.

[13] Alickovic, Emina&Subasi, Abdulhamit. (2011). Data Mining Techniques for Medical Data Classification.

[14] Maojo V., Sanandrés J. (2000) A Survey of Data Mining Techniques. In: Brause R.W., Hanisch (eds) Medical Data Analysis. ISMDA 2000. Lecture Notes in Computer Science, vol 1933. Springer, Berlin, Heidelberg.

[15] U. Fidan, İ. Sarı and R. K. Kumrular, "Classification of skin lesions using ANN," in *Medical Technologies National Congress (TIPTEKNO) 2016*, 2016, pp. 1-4.

CLASSIFICATION OF SKIN CANCER DISEASE USING DATA MINING TECHNIQUES

ORIGINALITY REPORT

20%	%	%	20%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	7%
2	Submitted to University of Hong Kong Student Paper	2%
3	Submitted to Middlesex University Student Paper	1%
4	Submitted to University of Mauritius Student Paper	1%
5	Submitted to Institute of Technology Blanchardstown Student Paper	1%
6	Submitted to Erciyes Ãniversitesi Student Paper	1%
7	Submitted to British University in Egypt Student Paper	1%
8	Submitted to Pondicherry University Student Paper	1%

9	Submitted to SRM University Student Paper	1%
10	Submitted to University of Southampton Student Paper	1%
11	Submitted to Southern New Hampshire University - Continuing Education Student Paper	1%
12	Submitted to Bridgepoint Education Student Paper	1%
13	Submitted to Amity University Student Paper	<1%
14	Submitted to University of Southern Queensland Student Paper	<1%
15	Submitted to University of Nizwa Student Paper	<1%
16	Submitted to University of South Alabama Student Paper	<1%
17	Submitted to 54972 Student Paper	<1%
18	Submitted to City University of Hong Kong Student Paper	<1%
19	Submitted to Higher Education Commission Pakistan	<1%