AN ENSEMBLE LEARNING APPROACH FOR THE PREDICTION OF ERYTHEMATO SQUAMOUS DISEASE

BY

ELIJAH TEMILADE QUEENSLY

PSC1808813

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE, FACULTY OF PHYSICAL SCIENCES, UNIVERSITY OF BENIN, BENIN CITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR AWARD OF BACHELOR OF SCIENCE (B.Sc.) IN COMPUTER SCIENCE

DECEMBER, 2022.

CERTIFICATION

This is to certify that this project work was carried out by ELIJAH TEMILADE QUEENSLY with matriculation number PSC1808813 in Computer Science, University of Benin, Benin City.

_____ _____

MR. E. C. IGODAN DATE

(PROJECT SUPERVISOR)

APPROVAL

This project is hereby approved in partial fulfillment of the requirement for the award of Bachelor of Science (B.Sc.) Degree in Computer Science at the University of Benin. Benin City.

_____ _____

MR. E. C. IGODAN DATE

(PROJECT SUPERVISOR)

_____ _____

PROF. A. O. EGWALI DATE

(HEAD OF DEPARTMENT)

DEDICATION

I dedicate this work to God, for giving me the strength and guidance to properly carry out and complete the work and also for his protection throughout my time at the University of Benin. This work is also dedicated to my parents and siblings, for their love, care, and support all through my academic pursuit. May the Almighty God bless all of them.

Lastly, this work is dedicated to my project supervisor. This work couldn't have been completed without you sir, GOD bless you.

ACKNOWLEDGEMENT

First and foremost, I want to express my gratitude to Almighty God for giving me the strength and resources necessary to complete this project.

I would also like to acknowledge my supervisor, Mr. E. C. Igodan, who put of his busy schedule took out time to proofread and put a finishing touch to this work, deserves my heartfelt thanks for his guidance, patience, knowledge, and unwavering support during this project.

It is with glowing passion that I record my highest admiration and greatest gratitude to the lectures of the Computer Science Department: Prof. (Mrs.) A. O. Egwali, Prof. (Mrs.) V. V. N. Akwukwuma, Prof. (Mrs.) F. A. Egbokhare, Prof. G. O. Ekuobase, Prof. F. Amadin, Prof. A. Imiavan, Dr. F. Chete, Dr. (Mrs.) V. I. Osunbor, Mr. S. O. P. Oliomogbe, Mr. E. E. Obasohan, Dr. (Mrs.) R. O. Osasere, Mrs. Linda Usiosefe, and many more.

Special thanks also to my best of friends: Egbobawaye Joseph, Mosiali Nneka Margaret, and Ofoluwa Oghenetega who in their own individual ways have imparted my life too with their words of knowledge/thoughts all throughout my undergraduate days in the university, I am very grateful.

I owe everything to my parents, Mr. and Mrs. Elijah Akintan and sisters; there is no doubt in my mind that they are the most important people in my life. I could never have gotten this far without their constant support and advice.

Finally, I would want to express my gratitude to every one of my classmates. Thank you all for all of your help during my time at the university.

TABLE OF CONTENTS

ABSTRACT

The global threat of cancer to human health is significant. It is one of the main factors contributing to rising mortality and morbidity, particularly in developing African countries. The lack of medical professionals in Nigeria's development of medical health centers is one of a number of issues related to cancer. Additionally, the fatality rate from most cancers, including skin cancer, has increased due to a lack of awareness, making treatment more challenging.

The research suggests using machine learning approaches for erythemato squamous disease diagnosis. The objective of the work is to develop and implement a hybrid ensemble approach using filter-embedded feature selection and ensemble of classifiers through bagging, boosting, and stacking.

The classification algorithms adopted in this study includes decision trees, naive bayes, multilayer perceptrons, support vector machines, KNN, and logical regression. While some of the feature selection methods are chi-square, information gain, reliefF, gain ratio, and recursive feature selection - SVM.

The performance of our models showed improved accuracy especially when using the ensemble methods. This project proves that using ensemble methods to predict erythemato squamous disease can address some of our challenges and the project also shows future prospects.

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF STUDY

Cancer, according to the national cancer institute, is a disease caused when cells divide uncontrollably and spread in surrounding tissues. The word cancer came from the father of medicine: Hippocrates, a Greek physician. Cancer can cause death if not diagnosed and treated in a timely fashion and can start almost anywhere in the human body, the skin inclusive (Kausar *et al.,* 2021). The skin is the body's largest organ made of water, proteins, fats and minerals. The skin protects the body from germs and regulates body temperature. They are three layers of the skin: epidermis, dermis and hypodermis. One inch of the skin has approximately 19 million skin cells and 60,000 melanocytes (Cells that make Melanie or skin pigments). It also contains 1,000 nerve endings and 20 blood vessels. Skin disease is the most common problem between people. Due to pollution and deployment of ozone layer, harmful UltraViolet (UV) rays of sun burn the skin and develop various types of skin diseases (Verma and Pal., 2019). From the investigation done by the skin cancer foundation, Skin cancer is the most common cancer on the global scale, where there are 9,500 cases of each diagnosed every day and an average of 2 people die from it every single hour in the United States (Jiang., 2021). In 2019 the global incidence of skin cancer was estimated to be over 104,350 cases, with almost 11,650 deaths (Ain *et al.,* 2020).

Basically, three types of skin disease appear on human skin: viral type, fungal type, and allergic type. The fungal and allergic disease type, can be cured if it is diagnosed properly and recognized at its early stage. But for the viral types, it is very necessary to identify the disease in its early stage (Bose *et al*., 2021 and Igodan *et al*., 2022).

Skin cancers often don't cause bothersome symptoms until they have grown quite large, then they may itch, bleed or even hurt. But typically they can be seen or felt long before they reach this point. Any unusual sore, lump, blemish, marking or change in the way an area of the skin looks or feels may be a sign of skin cancer or a warning that it might occur. The area might become red, swollen, scaly, crusty or begin oozing and bleeding. Symptoms of skin cancer vary depending upon the location of the cancer. Stage 1 melanoma is no more than 1.0 millimeter

thick which is about the size of a sharpened pencil point with or without an ulceration (broken skin).

There are three main types of skin cancer 1) basal cell carcinoma (BCC): This is the most common type of skin cancer, it looks like a flesh colored, pearl-like bump or pinkish patch of skin. 2) squamous cell carcinoma (SCC): The second most common type of skin cancer often looks like a red bump, scaly patch, or a sore that heals and then reopens. 3) melanoma: The deadliest form of skin cancer, frequently develops in the mole or suddenly appears as a new dark spot on the skin. Other types of skin cancer include: cutaneous T-cell lymphoma, dermatofibrosarcoma protuberans (DFSP), merkel cell carcinoma and sebaceous carcinoma.

Erythemato-squamous disease is also called skin disease. Skin disease can be classified into six classes cl:psoriasis, c2:seborrheic dermatitis, c3:lichen planus, c4:pityriasis roses, c5:chronic dermatitis, c6:pityriasis rubra. Skin disease diagnosis is difficult because disease classes possess identical clinical properties with very small changes. Basically, biopsy is used for the treatment of this skin disease (Verma *et al.,* 2020) and (Igodan *et al*. 2022).

Machine learning algorithms are widely used in medicine. Various disease diagnosis classification algorithms have been developed to provide high accuracy for predicting disease. Many machine learning algorithms are developed for predicting various types of disease at early stage after examining the various attributes of the disease. These algorithms are widely applicable in breast cancer, kidney diseases, thyroid diseases, diabetes, other cancer, erythemato-squamous diseases and many more (Igodan *et al.* 2022).

Ensemble learning as part of machine learning algorithms combines multiple learners, trained to solve similar problem and their predictions are combined with a single output that probably has better performance on average than any individual ensemble member. The idea behind ensemble learning is to combine weak learners into one strong learner, who has better generalization error and is less sensitive to overfitting in the presence of noise or small sample size. The four fundamental components of an ensemble learning are - training set, base inducer, diversity generator, and combiner (Tuysuzoglu *et al*., 2018).

In this research paper we selected Erythemato-squamous diseases for analysis. Various classification algorithms are applied and then ensemble methods are applied in this study. The

feature selection approach is applied with the classification algorithms to obtain classification accuracy for predictions (Verma *et al.,* 2019).

Motivated by (Verma and Pal., 2019) we adopted some classification algorithms in machine learning such as Logical Regression (LR), Naive Bayes (NB), K-nearest Neighbors (SVM), Decision Tree and so many more. Also some ensemble methods to enable us build a robust skin disease model for high classification performance such as bagging boosting and stacking was introduced.

Lastly to increase the accuracy of the prediction in the study we adopted another approach using feature selection such as filter and embedded methods.

1.2 RESEARCH MOTIVATION

Ensemble method is a powerful machine learning algorithm that is used across industries by data science experts .The beauty of the ensemble learning technique is that they combine predictions of multiple machine classifiers and produce better results with higher accuracy than using a single classifier.

One of the challenging tasks in skin diseases is due to the six classes in skin disease which possesses identical clinical properties with very small changes making diagnosis difficult and also from the literature researchers use only a few classification algorithms instead of ensemble methods.

Erythemato-squamous disease is a difficult problem in dermatology as these diseases display more than 90%, common features for both clinical and histopathological features. Motivated by these challenges, this project study applied feature selection, both the filter and embedded methods to handle feature relevance and redundancy, classical machine learning algorithms and their ensemble approaches using the bagging, boosting, and stacking methods so as to improve the classification performance of the models.

1.3 RESEARCH AIM AND OBJECTIVES

The aim of this project work is to develop a predictive model for the classification of Erythemato-squamous disease using Machine learning techniques

The specific objectives of this work are to;

a.) extract relevant features from erythemato-squamous disease dataset using ensemble feature selection techniques.

b.) design an ensemble classifier model for classification of (a) above.

c.) simulate (b) above using python programming language.

d.) evaluate the performance of (c) above using standard metrics.

1.4 RESEARCH METHODOLOGY

Machine learning is a technique for developing new algorithms which provides computer with the capability to learn from previously stored information. Table 1.1 shows a summary of the description of the dataset adopted in this study in terms of number of features, instances, attributes, type of data set, and other details.

The dataset is taken from the UCI Machine Learning Respiratory (http://archive.ics.uci.edu/ml).

The dataset contains 34 attributes, 33 of which are linear valued and one of them is nominal. There are six classes of Erythemato-squamous disease, with 366 instances and 34 features. The dataset is split into training and test set with 80% and 20% respectively.

Several combination of feature selection technique will be used to obtain optimal feature subsets. The goal of any feature selection method is to identify good features having maximum information with respect to the class labels of the sample instances.

After the feature selectors, the dataset are then standardized/normalized using the min-max model as represented in equation 1.1.

$$x^1 = \frac{x - xmax}{xmax - xmin} \qquad \text{..... (1.1)}$$

Then the subsets go through some pre-processing phases. Afterwards, several machine learning classifiers will be applied. Lastly, the ensemble method will combine the multiple classifiers to generate a strong model to improve the accuracy performance for the classification of skin disease datasets when fed with unknown datasets.

The classification algorithms adopted in this study includes decision trees, naive bayes, multilayer perceptrons, support vector machines, KNN, and logical regression. While some of the feature selection methods are chi-square, information gain, reliefF, gain ratio, and recursive feature selection - SVM

Table 1.1: Description of the dataset adopted (source:http://archive.ics.uci.edu/ml)

Data Set Characteristics:	Multivariate	Number of Instances:	366	Area:	Life
Attribute Characteristics:	Categorical, Integer	Number of Attributes:	33	Date Donated:	1998-01-01
Associated Tasks:	Classification	Missing Values?	YES	Number of Web Hits:	266889

1.5 SCOPE OF STUDY

The study focuses on ensemble of classifiers, features selection techniques and skin diseases (Erythemato-squamous disease).

1.6 SIGNIFICANCE OF STUDY

These projects, when completed, will serve as a second opinion for medical experts in the detection of skin diseases (Erythemato-squamous diseases).

It can also be adopted and adapted to any medical environment to meet specific needs for patient's care.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview

This chapter presents the knowledge of the general concepts about the human skin and skin cancers. Later in the chapter, the related work is presented for the classification, and ensemble of erythemato-squamous disease using the machine learning approach. The related work is present in tabular form. As per the discussion in the previous section, the use of machine learning algorithms in skin diseases for diagnosis purpose is not groundless. In this section, the literature is fully discussed. This section demonstrates in whole feature section, classification and ensemble of erythemato-squamous disease.

The rest of the chapter is organized as follows: in section 2.2, an introduction of the basics about the skin and skin cancer (symptoms, caused, types) followed by an interconnection of learning and radiology is presented. In section 2.3-2.6, the literature review of machine learning in radiology, as well as feature selection, and classification methods are presented, the related works using ensemble in skin disease classification is explored in section 2.7.

Finally, in section 2.8 an overall summary of the chapter is presented.

2.2 Introduction to skin cancer

The human skin is the outer covering of the body and is the largest organ of the integumentary system. The skin has up to seven layers of ectodermal tissue and guards the underlying muscles, bones ligaments and internal organs.

2.2.1 Basics of the Human Skin

The skin plays an important immunity role in protecting the body against pathogens and excessive water loss. It's other functions are insulation, temperature regulation, sensation, synthesis of vitamin D, and the protection of vitamin B folates.

In terms of surface area, the skin is the second largest organ in the human body. For the average adult human, the skin has a surface area of from 1.5-2.0 square meters (16-22 sqft). The thickness of the skin varies considerably over all part of the body.

The skin composed of three primary layers: the epiderm, the dermis and the hypodermis as shown in fig 2.1

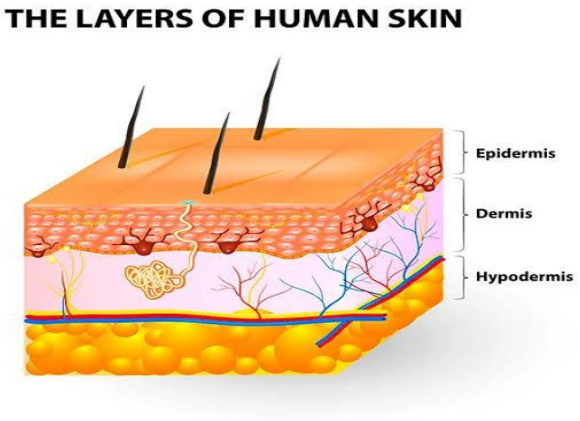


Fig 2.1: Layers of the human skin (https://www.oncobeta.com).

A) Epidermis is the outmost layer of the skin. It forms the water proof, protective wrap over the body's surface which also serves as a barrier to infection and is made up of satisfied squamous epithelium with an underlying basal lamina. The main type of cell that make up the epidermis are merkel cells, karatinocytes with melanocytes and langerhans cell also present.

The epidermis can be subdivided into the following: strata (beginning with the outermost layers), corneum, lucidum (only in palms of hands and bottoms of feet), granulosum, spinosum, basale.

B) Dermis is the layer of skin breath the epidermis that consists of cone connective tissue and cushions the body from stress and strain. The dermis is tightly connected to the epidermis by a basement membrane. It also harbors many nerve endings that provide the sense of touch and heat. It contains the hair follicles, sweet gland sebaceous glands, aprocrine glands, lymphatic vessels and blood vessels. The blood vessels in the dermis provide nourishment and waste removal from it's own cells as well as from the stratum basale of the epidermis.

The dermis is structurally divided into two areas: a superficial area adjacent to the epidermis, called the papillary region and a deep thicker area known as the reticular region.

C) Hypodermis (also subcutaneous tissue and subcutis) is not part of the skin but lies below the dermis of the cutis. It's purposes is to attach the skin to underlying bone and muscle as well as supplying it with blood vessels and nerves. It consists of loose connective tissue, adipose tissue and elastin. The main cell types are fibroblasts, macrophages and adipocytes (subcutaneous tissue contains to 50% of body fat) (en.m.wikipedia.org).

2.2.2 Skin cancer

Skin cancer is the abnormal growth of skin cells. if not treated, can grow larger and cause problems beneath the skin, sometimes damaging the muscles and bones (www.uofmhealth.org).

Found early, skin cancer is highly treatable. Often a dermatologist can treat an early skin cancer by removing the cancer and a bit of normal-looking skin. Given time to grow, treatment for skin cancer becomes more difficult (waad.orgww). Skin cancer rates are higher in woman than in man before age 50, but are higher in men after age 50, which may be related to different in recreation and work related UV exposure (www.aad.org).

To determine a skin cancer's stage or severity, your doctor will factor in how large the tumor is it has spread to your lymph node and if it has spread to other parts of the body. Skin cancers are divided into too primary groups for staging purposes: non melanoma skin cancer and melanoma (www.healthline.com).

Non melanoma skin cancers include basal cell and squamous cell cancers.

a. Stage O: The abnormal cells have not spread beyond the out most layer of skin, epidermis.

b. Stage I: The cancer may have spread to the next layer of skin the dermis, but it is no longer than too centimeters.

c. Stage II: The tumor is longer than two centimeters, but has not spread to nearby sites or lymph nodes.

d. Stage III: The cancer has spread from primary tumor to nearby tissue or bone, and it is larger than three centimeters.

e. Stage IV: The cancer has spread beyond the primary tumor site to nodes and done or tissue. The tumor is also larger than three centimeter.

Melanoma stages include:

a. Stage O: This noninvasive type of skin can has not penetrated below the epidermis.

b. Stage I: The cancer may have spread to the second layer of skin, the dermis but it remains small.

c. Stage II: The cancer has not spread beyond the original tumor site, but it is larger thicker, and may have other signs and symptoms. These include scaling, bleeding, or flaking.

d. Stage III: The cancer has spread or metastasized to your lymph nodes or to the nearby skin or tissue.

e. Stage IV: The most advanced stage of melanoma, stage IV is an indicator the cancer has spread beyond the primary tumor and is showing up in lymph nodes, organs or tissue distant from the original site.

When cancer comes back after treatment, it's called recurrent skin cancer. Anyone who has been diagnosed with and treated for skin cancer is at risk for a recurrence of the cancer.

That makes follow up care and self-examination even more important (www.healthline.com).

2.2.3 Causes of skin cancer

It generally develops in areas that are exposed to sun, but it can also form in places that don't normally get sun exposure. Both types of skin cancer occur when mutations develop in the DNA of your skin cells. These mutations cause skin cells to grow uncontrollably and form a mass of cancer cells (www.healthline.com).

Basal cell skin cancer is caused by ultraviolent (UV) rays from the sun or tanning beds. UV rays can damage the DNA inside the skin cells, causing the unusual cell growth. Squamous cell skin cancer is also caused by UV exposure (www.healthline.com).

Squamous cell skin cancer can also develop after long term exposure to cancer –causing chemicals. It can develop within a burn scar or ulcers, and may also be caused by some types of human papillomavirus (HPV).

There is strong evidence that: drinking water contaminated with arsenic increases the .risk of malignant melanoma. There are also some evident that greater birth weight might increase the risk of malignant melanoma, also consuming alcoholic drinks might increase the risk of malignant melanoma and basal Carcinoma (www.wcrf.org).

Other causes of skin cancers include: (over exposure to certain types of light, sun as ultra-violet rays from the sun or tanning devices), medication (medicines used to suppress the immune system after organ transplantation), infection (infection with human papilloma virus), Genetics and family history (some rare mutation in specific genes can lead to skin cancer. Having a family history of skin cancer also increase risk), skin pigmentation (skin cancer is more common in lighter-skinned populations) (www.wcrf.org).

2.2.4 Symptoms of skin cancer

Skin cancers aren't all identical, and they may not cause may symptoms. Skill, unusual changes to your skin can be a wearing sign for the different type's cancer including:

a. Skin lesions: A new mole, unusual growth, bumps, sore. Scaly path, or dark spot develops and doesn't go away.

b. Asymmetry: The two halves of the lesion or note aren't even or identical.

c. Border: The lesions have ragged, uneven edges.

d. Color: The spot has an unusual color, such as white, pink, black, blue, or red.

e. Diameter: The spot is larger than one-quarter inch or about the size of a pencil eraser.

f. Evolving: you can detect that the mole is changing size, color or shape (www.healthline.com).

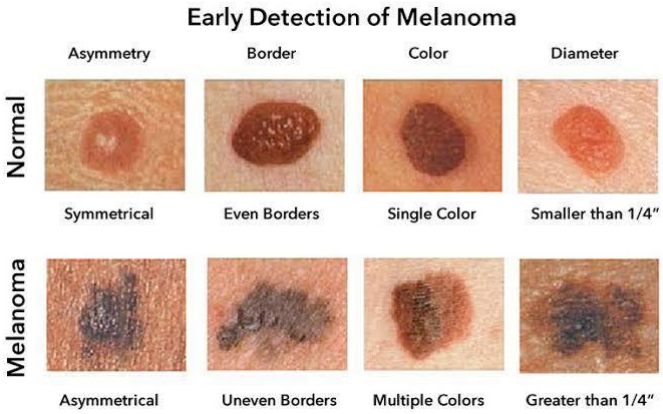


Fig 2.2: Types and shapes of the diseases (source: https://encrypted-tbn0.gstatic.com)

Basal cell carcinoma may appear as: A pearly or way bump or a flat, flesh-colored or brown scar-like lesion, or a bleeding or scabbing sore that heals and returns. While squamous cell carcinoma may appear as: A firm, red noddle, or a flat lesion with a scaly, crusted surface. While melanoma signs include: A large brownish spot with darker speckles, or a mole that changes in color, size or feel or that bleeds, or a small lesion with an irregular border and portions that appear red, pink, white, blue or blue-black, or a painful lesion that itches or burns, or dark lesions on your palms, sales, finger tips or toes or on mucous membranes giving your, mouth, nose, vagina or anus (www.mayoclinic.org).

2.2.5 Types of Skin Cancer

The type of skin cancer a person gets is determined by where then cancer begins. If the cancer begins in skin cells called basal cells, the person has basal cell skin cancer. When cells that give air skin its color become cancerous, melanoma develops (www.aad.org)

There are 4 main types of skin cancer:

1. Basal cell carcinoma: Basal cells are the round cells found in the lower epidermis. About 80% of skin cancers develop from this type of cell. These cancers are described as basal cell carcinomas. Basal cell carcinomas most often develops on the head and

neck, although it can be found anywhere on the skin. It is mainly caused by sun exposure or develops in people who received radiation therapy as children. This type of skin cancer usually grows slowly and rarely spreads to other parts of the body.

2. Squamous cell carcinomas: Most of the epidermis is made up of flat, scale-like cells called squamous cells. Around 20 % of skin cancers develop from these cells, and these cancers are called squamous cell carcinomas. Squamous cell carcinomas are mainly caused by sun exposure, so it may be diagnosed on may regions of the skin. It can also develop on skin that has been burned, damaged by chemicals, or exposed to x rays. Squamous cell carcinoma is commonly found on the lips at sites of a long-standing scar, and on the skin outside the mouth, anus, and a woman's virginal. About 2% to 5% of squamous cell carcinomas spread to other parts of the body.

3. Markel cell cancer: Markel cell cancer is a highly aggressive, or fast-growing, rare cancer. It starts in hormone-producing cells just beneath the skin and in the hair follicles. It is usually found in the head and neck region. Markel cell cancer may also be called neuroendocrine carcinomas of the skin.

4. Melanoma: There are scattered cells called melanocytes where the epidermis meets the dermis. These cells produce the pigment melanin which gives skin its color. Melanoma starts in melanocytes, and it is the most serious type of skin cancers.

Basal cell carcinoma and squamous cell carcinoma are sometimes grouped together and called keratinocyte carcinoma. This is because they begin in a type of skin cell called a keratinocyte.

These are a few other, rare types of skin cancer, including cutaneous (skin) lymphomas, Kaposi sarcoma, skin adnexal tumors and sarcomas, all which are classified as non –melanoma skin cancers (www.cancer.net).

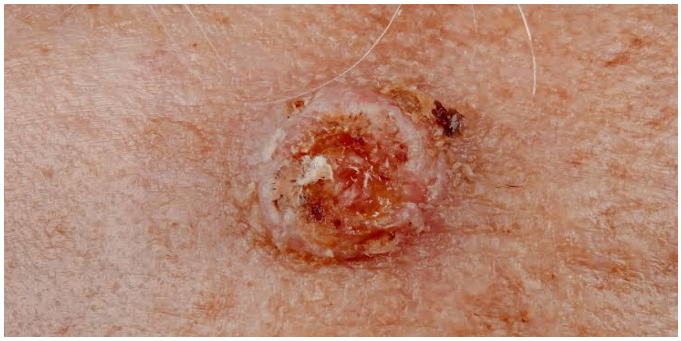


Fig2.3 squamous-cell-carcinoma (source: https://encrypted-tbn0.gstatic.com)

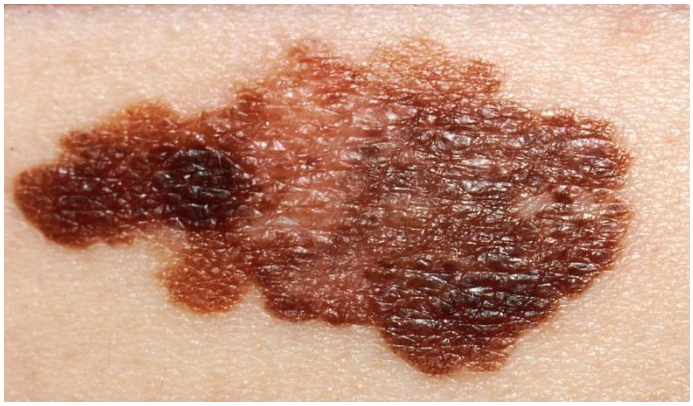


Fig 2.4 melanoma (source: https://upload.wikimedia.org)

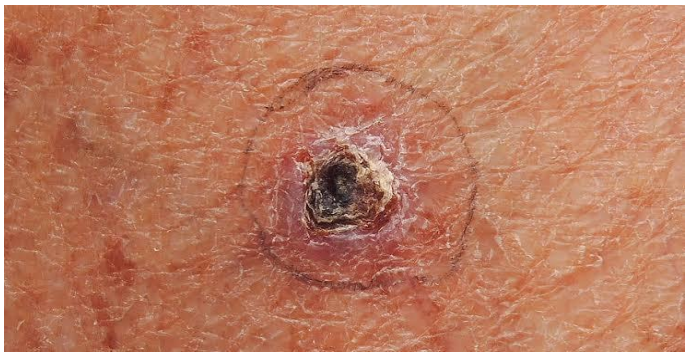


Fig 2.5 basal-cell-carcinoma (source: https://encrypted-tbn0.gstatic.com)

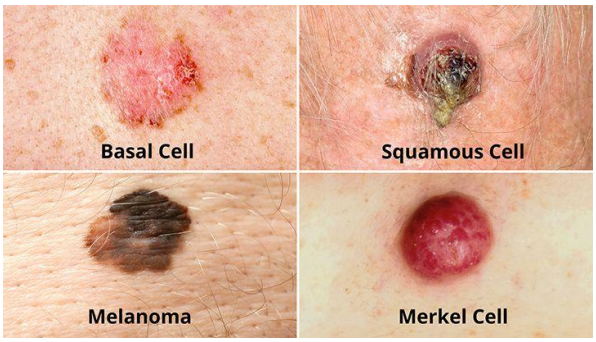


Fig 2.6 Different types of skin cancer (source: https://www.everydayhealth.com)

2.2.6 Skin Cancer Diagnosis

Treatment depends upon the stage of cancer stages of skin cancer ranges from stage O to stage IV. The higher the number, the more cancer has spread (www.my.cleverlandclinic.org).

Due to the complexity of human skin texture and the visual closeness effect of the diseases, sometimes it is really difficult to detect the exact type. Therefore, it is necessary to detect and recognize the skin disease at its very first observation (Bose *et al*., 2021).

Computational diagnosis systems have been developed to assist dermatologists in early diagnosis of skin cancer from dermoscopic images (Oliveira *et al*., 2017).

To diagnose skin lesions, dermatologists have been using an advanced imaging device, namely dermoscopy, to improve the clinical analysis of malignant melanoma. Dermoscopy is a noninvasive diagnostic device that is commonly used in clinical diagnosis, and it is useful for the early detection of melanoma unfortunately, dermoscopy is costly, often not available, and has limited use in some countries such as USA, Taiwan, and India. Furthermore, the proposed approaches based on dermoscopic images are limited to professional dermatology settings. Indeed, it is unlikely that general practitioners would have such a specialized device. Indeed, May dermatologists and general practitioners use optical skin lesion images acquired by a standard Carrera.

In addition, due to improvement in image processing methods in recent years, and owing to the recurrent patterns in skin lesions, there has been an important increase in interest in the

development of CAD systems for melanoma detection. Hence, a CAD system can efficiently improve accuracy diagnosis, taking advantages of the imaging techniques and decision making methods (Verma and Pal, 2019).

The types of erythemato-squamous diseases (ESD) are given in Table2.1. These diseases are frequently noticed in white skinned people in USA,UK and Europe; whereas these diseases are reported very less in Asian countries. As these six diseases display similar clinical features such as erythema and scaling with minor variations, the detected of erythemato-squamous disease (DESD) is complex. And another added problem is, a disease showing the features and signs of another disease at initial phase, may display its own characteristics features at following stages (Badrinath*et al*., 2015).

At the initial stages, these diseases seem similar with the erythema and scaling. When analyzed in depth, some of the patents show the typical chemical characteristics of the diseases at the predilection sites while other group of patients shows typical localizations. The chemical evaluation of patients was performed with 12 features that are classified as chemical features. In differential diagnosis, the degree of erythema and scaling, the presence of itching, whether the oral mucosa, whether the border of lesions are definite or not, presence of koebner phenomenon. The formation of papules, knees, scalp and elbows are involved or not are considered and taken into account. The family is one of features considered during the analysis. By analysis, the koebner phenomenon is present only in pityriasisrosea, lichen planus and psoriasis. The erythema and scaling features of psoriasis is greater than chronic dermatitis. Further, itching and follicular papules are for pityriasisrubra pilaris, whereas polygonal papules are for lichen planus. The predilection site for lichen planus is oral mucosa, whereas elbow, knee and scalp involvement are for psoriasis. Family history as normally present for pityriasisrubra pilaris and psoriasis. It generally starts during very early age, in childhood (Badrinath*et al*., 2015).

Some of the patients are diagnosed with clinical features, however for a definite diagnosis, a biopsy is generally done. Samples of skin were evaluated for 22 histopathological features. Another complex for differential diagnosis is, a disease showing the histopathological features and signs ofanother disease at initial phase, may display its own characteristic features in the next following stages. Some of the skin samples display the typical histopathological features of the disease whereas others do not. A diagnostics feature for lichen planus is melanin

incontinence exocytosis may be seen in seboreic dermatitis, lichen planus and pityriasisrosea. A diagnostic feature for chronic dermatitis is fibrosis of the papillary dermus. At different levels of these disease acauthosis and parakeratosis are seen. The diagnosis features of psoriasis are thinning of the supra papillary epidermis and clubbing of the rete ridges. The diagnostic features of lichen planus are saw tooth appearance of rete's, disappearance of the granular layer, a band like infiltrate, vacuolization and damage of basal layer. The diagnostic features for pityriasisrubra pilaris are follicular horn plug and perifollicular parakeratosis (Badrinath *et al*., 2015).

Sometimes a biopsy alone can remove all the cancer tissue if the cancer is small and limited to the skin's surface only. Other common skin treatments, used alone or in combination, include:

Cryotherapy : uses liquid nitrogen to freeze skin cancer, the dead cells slough off after treatment.

Excisional surgery: involves removing the tumor and some surrounding healthy skin be sure all cancer gas been removed.

5. Molis surgery: the visible, reused area of the tumor is removed first. Then your surgeon uses a scalpel to remove a thin layer of skin cancer cells. The layers are examined under a microscope immediately after removed. Additional layers of tissue continue to be removed, the layer at a time, until no more seen under the microscope.

6. Curettage and electrodessication: this technique uses an instrument with a sharp loped edge to remove cancer cells as it scrapes across the tumor. The area is then treated with an electric needle to destroy and remaining cancer cells.

7. Chemotherapy and immunotherapy: chemotherapy uses medication to kill cancer cells. Anticancer medications can be can be applied directly on the skin(topical chemotherapy) if limited to your skin's top layer or provided through pills or an iv if cancer has spread to other parts of your body. Immunotherapy uses your own body's immune system to kill cancer cells.

8. Radiation Therapy: uses radiation (strong beams of energy) to kill cancer cells or keep them from growing and dividing.

9. Photodynamic therapy: your skin is coated with mediation and a blue red fluorescent light then activates the medication. Photodynamic therapy destroys precancerous cells while leaving normal cells while leaving normal cells alone. (www.my.clevelandclinic.org)

Table2.1 : Dataset used – 34 features and diseases (Badrinath*et al.,* 2015).

Diseases	**Clinical features**	**Histopathology features**
1. Psoriasis	1. Erythema	1. Melanin incontinence
2. Seboric dermatitis	2. Scaling	2. Eosinophils in the infiltrate
3. lichen planus	3. Definite borders	3. PNL infiltrate
4. Pityriasis rosea	4. Itching	4. Fibrosis of the papillary dermis
5. Chronic dermatitis	5. Koebner phenomenon	5. Exocytosis
6. Pityriasis rubra pilaris	F6 Polygonal papules	6. Acanthosis
	7. Follicular papules	7. Hyperkeratosis
	8.Oral mucosal involvement	8. Parakeratosis
	9. Knee and elbow	9. Clubbing of the rete ridges involvement
	10. Scalp involvement	10. Elongation of the rete ridges
	11. Family history	11. Thinning of the suprapapillary
	12. Age	12. Pongiform pustule epidermis
		13. Munro microabcess
		14. Focal hypergranulosis
		15. Disappearance of the granular layer
		16. Vacuolization and damage of basal layer
		17. Spongiosis
		18. Saw-tooth appearance of retes
		19. Follicular horn plug
		20. Perifollicular parakeratosis
		21. Inflammatory mononuclear infiltrate
		22. Band-like infiltrate

2.3 Machine learning

A machine learning Algorithm is the by which the AI system conducts it's task generally predicting output values from given input data. In reality, machine learning is about setting systems to the task of searching through data to look for patterns and adjusting actions accordingly (www.recordedfuture.com) we can also say machine learning is the technique for developing new algorithm which provides computer the capability to learn from previously stored information (verma *et al*., 2019).

Machine learning is classified into three broad domains i.e supervised machine learning, unsupervised machine learning and reinforcement machine.

2.4 Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and in some cases, to improve the performance of the model (www.machinelearningmastery.com)

Three benefits of performing feature selection before modelling the dataset are:

1. Reduces the model complexity
2. Reduces over fitting
3. Improves Accuracy
4. Reduces training time

We can obtain the feature importance of each feature of dataset by using the feature importance property of the model feature importance provides a score for each for each feature of dataset the higher the score the more important or relevant is the feature towards output variable (verma *et al*., 2019).

The simple rule is garbage in then we find garbage out. This means if we find unnecessary attributes to a classified we get undesired results. Important features are those features which provide efficient predictions (verma *et al*., 2017).

The feature selection process aim to find the best feature subsets to generate the ensembles of classifiers, feature selection algorithms are usually a combination of both search and evaluation methods (oliveria *et al*., 2017).

There are many methods of selecting features. The feature selection methods are broken down into three basic categories: Filters, wrappers and embedded methods In recent years, researchers have developed many methods to select features through IT tools still new feature selection methods are being proposed.

The rapidly increasing number of features is a very serious problem to be solved. This increases the computational complexity of the algorithm, extends the learning process and increases multi-level classification method.

The best result of the classifies is given by a property selected feature selection algorithm. Feature selection reduction of the feature space dimensionally, reduces the number of free parameters in the classifies necessary for estimation. When collecting data again, we can focus only on the feature important for the classification algorithm. Filters mainly use the general characteristics of data sets. Wrappers and embedded methods build a subset of functions based on selected algorithm.

The most important algorithm for selecting the features of medical images include methods: SBS (sequential backward selection), SFSC (sequential forward selection) and its modifications SFFS (sequential forward floating search).

Other methods are method Plus-L-Minus-R, NNFF (Nearest neighbor with feature projection), methods based on genetic algorithms, OSA(Oscillating Search Algorithm), methods based on the use of fractal dimension, methods based on information they (Michalska., 2021).

A. Filter methods

Filter methods pick up the intrinsic proprieties of the feature measured via uninvariate statistics instead of cross-validation performance. These methods are faster and less computationally expensive them wrapper methods. When dealing with high dimensional data, it is computationally cheaper to use filter methods. (www.analyticsvidhya.com) The most frequently

used by researchers are filter methods. The methods uses statistical measure and the functions are selected for retention or removal from the data (Michalska., 2021).

1. Chi-square: the Chi-square test is used for categorical features in dataset. We calculate chi-square between each feature and the target and select the desired number of features with the best chi-square scores. In order to correctly apply the chi-squared in order to test the relation between various features in the dataset and the target variable, the following conditions has to be met. The variables have to be categorical, sampled independently and values should have an expected frequency greater than 5.

2. Information Gain: Information gain calc7ulates the reduction in entropy from transformation of a dataset. It can be used for feature selection by evaluating the information gain of each variable in the context of the target variable.

3. Fisher's Score: Fisher score is one of the most widely used supervised feature selection methods. The algorithm which we use returns the ranks of the variables based on the fisher's score in descending order.

4. Correlation Coefficient: Correlation is a measure of the linear relationship of 2 or more variables. Through correlation, we can predict one variable from the other. The logic behind using correlation for feature selection is that the good variables are highly correlated with the target. Furthermore, variable should be correlated with the target but should be uncorrelated among themselves (www.analyticsvidlya.com).

B. Wrapper methods: Wrappers require some method to search the space of all possible subsets of features, assessing their quality by learning and evaluating a classifier with that feature subset. The feature selection process based on a specific machine learning algorithm that we are trying to fit on given dataset. It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion. The wrapper methods usually result in better predictive accuracy than filter methods. Let's, discuss some of its techniques.

    1. Forward Features Selection: This is an iterative method where start with the best performing variable against the target. Next, we select another variable that gives the

best performance in combination with the first selected variable. This process continues until the present criterion is achieved.

2. Backward Feature Elimination: This method works exactly opposite to the forward feature selection method. Here, we start with all the features available and build a model. Next, we select the variable from the model which gives the best evaluation measure value. This process is continued until the present criterion is achieved.

3. Exhaustive Feature Selection: This is the most robust feature selection method covered so far. This is a brute force evaluation of each feature subset. This means that it tries every possible combination of the variable and returns the best performing subset.

4. Recursive Feature Elimination: Given an external estimation that assigns weights to features, the goal is to select features by recursively considering smaller and smaller sets of features (www.analyticsvidhya.com).

C. Embedded Methods: These methods encompass the benefits of both the wrapper and the filter methods by including interactions of features but also maintaining reasonable computational cost. Embedded methods are iterative in the sense that takes care of each iteration of the model training process and carefully extracts those features which contributes the most to the training for a particular iteration. Let's discuss some of these techniques:

1. LASSO Regularization (L1): Regularization consists of adding a penalty to the different parameters of the machine learning model to reduce the freedom of the model i.e to avoid over-fitting in linear model regularization, the penalty is applied over the coefficients that multiply each of the predictors. From the different types of regularization, Lasso or L1 has the property that is able to shrink some of the coefficients to zero. Therefore, that feature can be removed from the model.

2. Random Forest Importance: Random forest is a kind of a bagging algorithm that aggregates a specified number of decision trees. The tree based strategies used by random forests naturally rank by how well they improve the purity of the node or in

other words a decrease in the impurity (Gini impurity) over all trees. Nodes with the greatest decrease in impurity happen at the start of the trees, while notes with the least decrease in impurity occur at the end of trees. Thus, by pruning trees below a particular node, we can create a subset of the most important features.

There is also a tendency to mix algorithms, as in the case of hybrid methods, which usually combine two or more feature selection algorithms of different conceptual origins sequentially. A typical example is to first apply a less computationally costly filter to remove some features and then use a more computationally costly wrapper for fine tuning. Most of the new features selection methods that appear are filters, although we can find representative methods for all three categories. A large number of feature selection methods now available therefore complicate the choice of the best method for a given problem.

2.5 Classification Algorithms

A classification algorithm, in general is a function that weighs the input features so that the output separates one class into positive values and the other into negative values.

Classification is the process of recognizing, understanding and grouping ideas and objects into present categories or "sub-population" using pre-categorized train9ig dataset, machine learning programs use a variety of algorithms to classify future datasets into categories.

Classification algorithm in machine learning use input training data to predict the likelihood that subsequent data will fall into one of the predetermined categories. One of the most common uses of classification is filtering emails into "spam" or "non-spam"

In short, classification is a form of "pattern recognition" with classification algorithms applied to the training data to find the same pattern (similar words or sentiments, number sequence etc) in future sets of data (www.monkeylearn.com).

The study of classification in statistic is vast and there are several types of classification algorithms you can use depending on the dataset you are working with. Below are seven of the most common algorithm in machine learning.

2.5.1 Logistics Regression (LR)

Logistics regression is a calculation used to predict a binary outcome either some happens, or does not. This can be exhibited as yes/no, pass/fail, alive/dead, etc. independent variables are analyzed to determine the binary outcome with the results falling into one of two categories. The independent variables can be categorical or numeric, but the dependent variable is always categorical, written like this:

$$PCY = \frac{1}{X} \quad OR \quad P(Y = \frac{0}{X})$$

It calculates the probability of dependent variables V, given independent variable X.

This can be used to calculate the probability of a word having a positive or negative connotation (o or in a state between). Or it can be used to determine the object contained in a photo (tree, flower, grass, etc.), with each object given a probability between 0 and 1 (www.monkeylearn.com).

2.5.2 Naïve Bayes (NB)

Naïve Bayes calculates the possibility of whether a data point belongs within a certain category or does not. (www.monkeylearn.com).

Naïve Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naïve Bayes classifies work well in many real world situations such as document classification and spam filtering.

This algorithm requires a small amount of training data to estimate the necessary pies are extremely fast compared to more sophisticated methods (www.analysticsindiamag.com).

2.5.3 K-nearest Neighbors (KNN)

KNN base learner is used for classification and regression problems but KNN is generally used in classifying problems. KNN is a lazy learning and non-parametric algorithm. If there is no assumption for underlying data distribution, then it is called non-parametric , it means the model structure determined from the data set. KNN will be helpful in prediction, where the data sets do not follow mathematical hypothetical assumptions. KNN does not need any training for data for

development of model, therefore, it is called lazy learning algorithm. All training data are used in directly testing phase (Verma *et al*., 2020).

KNN is also a supervised learning method. In this method, each sample classified it's surroundings samples. Therefore, when it is needed to classify an unknown sample, then it could be classified by its surroundings samples. When a training dataset and an unknown sample are given, then the classifier calculates the distance between the unknown sample and all datasets. The smallest distance value between the unknown sample and the training set is accepted. Primarily, the performance of KNN depends on the value of K, the no of nearest neighbor, to classify an unknown sample. If K is very small then the classification is not very good. But the largest number of K gives the perfect result of classification (Bose *et al.*, 2021).

2.5.4 Decision tree (DT)

Decision trees are a type of supervised machine learning. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features (verma *et al*., 2019).

A decision tree is a tree structure, where each interior node represent a test on a feature, each branch denotes a result of the test and each terminal node holds a class label (Verma *et al*., 2020).

It divides a dataset into various smaller data subset and this forms associated decision tree. The end result is a tree with decision nodes which has two or more branches and leaf nodes which represents a classification or decision. The decision node in the tree represents the best predictor called the root node. Decision tree can use categorical data and digital data (Verma and Pal., 2019).

2.5.5 Random Forest (RF)

The random forest algorithm is an expansion of decision tree, in that you first construct a multitude of decision trees with training data, then fit your new data within one of the trees as a "random forest" (www.monkeylearn.com).

Random forest classifier will handle the missing values. When we have more trees in the forest, random forest classifier won't over fit the model. (Verma *et al*., 2019).

This algorithm is most simple and flexible to use. Random forests randomly select data to create decision trees and give prediction from each tree and choose the best solution by use of voting technique. It also provides an attractive excellent display of the feature importance (Verma and Pal., 2019).

2.5.6 Support Vector Machines (SVMs)

It is a supervised learning algorithm. This algorithm transforms the complex data based on the kernel function. It maximizes the separation between the classes to make a clear prediction (Bose *et al*., 2021).Support vector machines are discriminant classifiers that are correctly defined by separate hyper planes. The SVM uses labeled training data (supervised learning) and the algorithm outputs the best hyper plane, which classifies the new record (Verma and Pal., 2019).Support vectors are simply vector machine is a frontier which best segregates the two classes (hyper-plane/line) (Verma *et al*., 2019).

2.5.7 Multilayer perception (MLP)

A multilayer perception is a logistic regression classifier. In this classifier input data is changed with the help of a learnt non-linear conversion. This change input data into a layer, where input data becomes linearly divisible.

This layer which changes data from input is called hidden layer. Only single hidden layer is used in multilayer perception otherwise it will work as ANN. Although the use of multiple hidden layer are very beneficial (Verma *et al*., 2020).

2.6 ENSEMBLE TECHNIQUES

The ensemble method is used as a method to find the accuracy of the skin disease dataset to improve the performance of algorithm (Verma *et al*., 2019). Ensemble methods are used for combining several base leanings to predict a problem enhancing the perdition of single classifiers. Ensemble methods can be broadly categorized into two types joining multiple classifiers of similar types and joining multiple classifiers of different types (Verma *et al*., 2020).

The three main classes of ensemble learning methods are bagging, stacking and boosting.

2.6.1 Bagging Ensemble Learning

Bootstrap aggregation, or bagging for short, is an ensemble learning method that seeks a diverse group of ensemble members by varying the training data (www.machinelearningmastery.com).

It's a machine learning model aggregation technique designed to improve the stability and accuracy and to reduce variance to avoid over fitting of machine learning algorithms applied in regression and classification methods (Verma *et al*., 2019).

2.6.2 Boosting Ensemble Learning

Boosting ensemble technique is used to create a set of predictors in this technique, learners learn t sequentially, while near the beginning learners fir simple model to the data and then analyze error in the data.

Continuous trees (random samples) are suitable and in each step, the goal is to improve the accuracy of the previous tree. When an input is misclassified by an assumption, its weight increases, so the next assumption is more likely to be correctly classified. This process turns base learners with weak capabilities into better performing models (Verma *et al*., 2020).

2.6.3 Stacking Ensemble Learning

Stacking ensemble method is used for combining multiple base learners of different type with the help of a meta-classifier. Six base learners NB, KNN, DT, SVM, RF and MLP are trained with whole training dataset, and then, meta-classifiers is applied on the outputs "meta-features" of each base learner. Mata-can be trained on the probabilities from the ensemble techniques or predicted class labels (Verma *et al*., 2020).

2.7 RELATED WORK

In this section, we review in a tabular form, the current state-of-art literature related to brain tumor classification and detection using ensemble machine learning techniques.

Numerous techniques have been reviewed for the classification of brain tumor using ensemble as shown in the Table 2.2:

Table 2.2: Literature Review

Author(s)/ Year	Classifiers/Machine Learning	Feature Selection Methods	Model Accuracy Achieved (%) Metrics	
1) Oliveira *et al.,* (2017)	Optimal Path Forest(OPF), Bagging, Ada Boost, Random Forest	Pearson's Correlation Coefficient, Gain Ratio, Information Gain, Relief-F, Principal-Component Analysis (PCA), Correlation-based Feature Selection (CFS)	ACC: 94.3 SEN: 91.8 SPE: 96.7	Lacking deep l High computat difficulty in an focus on homo using only OP Classifier) and segmentation p
2) Verma *et al.,* (2020)	Naïve Bayesian (NB), K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Multilayer Perception (MLP), Boosting, Bagging, Stacking	Hybrid, Chi-Square, Information Gain, Principle Component Analysis (PCA)	ACC: 99.67	Only filter bas adopted in this
3) Verma *et al.,* (2019)	Classification and Regression Trees (CART), Support Vector Machine (SVM), Random Forest, Decision Tree, Gradient Boosting Decision Tree (GBDT).	-	ACC: 98.64	No feature sele
4) Verma and Pal. (2019)	Naïve Bayesian (NB), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Stacking.	Univariate Feature Selection, Feature Importance and Correlation Matrix with Heatmap, Chi-square	ACC: 99.86	Increase time c FS methods in subsets by voti features.
5) Verma *et al.,* (2019)	Bagging, Ada Boost, Gradient Boosting, Passive Aggressive Classifier, Linear Discriminate Analysis, Radius Neighbours Classifiers, Bernoulli Naïve Bayesian, Gaussian Naïve Bayesian, Extra Tree Classifier.	Feature Importance Method	ACC: 99.68	Time complex features used a method.
6) Bose *et aL.* (2021)	Bagged Tree Ensemble, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Deep Neural	-	ACC: 92.99 SPE: 99.93	No Feature Sel

	Network, VGG16, GoogleNet, ResNet50		SEN: 98.68	
7) Verma *et al.,* (2019)	Bagging, Ada Boost, Gradient Boosting, Passive Aggressive Classifier (PAC), Linear Discriminant Analysis (LDA), Radius Neighbours Classifier (RNC), Bernoulli Naïve Bayesian (BNB), Gaussian Naïve Bayesian (NB), Extra Tree Classifier (ETC)	Feature Importance Method	ACC: 99.68	Few Feature S therefore ,Tim of feature used method
8) Badrinath *et al.,* (2013)	Fuzzy Extreme Learning Machine (FELM), ANN, ANFIS, SVM, AdaBoost, Hybrid AdaBoost, ELM ABC-FELM, Treshold-based ABC-FELM	Hybrid, Artificial Bee Colony (ABC)	ACC: 99.57	Time complex
9)Badrinath *et al.,* (2015)	AdaBoost, real, modest and gentle AdaBoost, and Hybrid AdaBoost, Fuzzy Logic, Neural Networks, SVM, ANFIS	-	ACC: 99.3	No Feature Se data over fittin
10) Michalska (2021)		Filter Methods, Wrappers Methods, Embedded Methods, Correlation based, Consistency-based filter, Information Gain, Relief-F, Pearson's Correlation, LDA (Linear Discriminant Analysis), ANOVA, Chi-square, Fisher Score.	-	No Ensemble/
11) Jiang (2021)	ResNet 50, ResNeXt 50, ResNeXt 101, EfficientNet-B4, MobileNetV2, MobileNetV3-Large, MnasNet, Majority Voting, Weighted Voting by Precision, Weighted Voting by Cubic Precision.	-	ACC: 76.71	Failed to show feature selecti complexity.
12) Kausar *et al.,* (2021)	ResNet, InceptionV3, DenseNet, Inception ResNetV2, VGG-19, Majority Voting, Weighted Majority Voting, Weighted Average.	-	ACC: 98.6	No Feature Se prolonged time
13) Rubeena *et al.,* (2020)	Gaussian Filter, Morphological Filter	-	-	No Feature Se

14) Abbes and Sellami (2021)	VGGNet16, ResNet18, and ResNet50	-	ACC: 97 SEN: 100 SPE: 94	Dataset is sma No Feature Se
15) Ain *et al.,* (2020)	Random Forest, Bagging, AdaBoost, LogitBoost, Random Committee SVM, k-NN, MLP, J48, NB	Embedded-GP, Wrapper-GP, Patino et al., Kawahara et al.,	ACC: 100	Low Construct investigation i
16) Igodan et al. (2022)	SVM, KNN, DT, naïve Bayes, MLP	Info gain, gain ratio, ReliefF, Chi squared RFE-SVM, PRIFEB, MIFEB	Acc: 92.9% SEN: 85.8% SPE: 97.4%	Lack of Comp study Lack of genera

2.8 SUMMARY OF CHAPTER

One of the leading causes of death worldwide is cancer, which is a disease in which the cells of the body grow uncontrollably and spread to other parts of the body; Cancer also develops when your genes lose control over how your cells divide. For example, instead of dying, old cells grow and form abnormal cells.

This chapter introduced the basic concepts about cancer, its structure, and some types of cancers. Some primary basics have discussed the types of cancers, their symptoms, and the possible causes through which the cancer spreads in the human body, followed by the use of functional genomics to identify the importance of genes for cancer diagnosis.

Further, an interconnection between the machine learning system and functional genomics is discussed. The chapter focuses on the various feature selection methodology and the classification model. The subsection discusses the various machine learning approaches that are used.

Later in the chapter, we presented most of the state-of-art techniques employed for the classification of different types of cancers in microarray-based data. The presented related work explored the machine learning system in regard to the identification and classification of cancers

using an ensemble in trying to distinguish genes that play a functional role in different types of cancer.

In the next chapter, the methodology is presented for cancer classification using an ensemble. The chapter gives a detailed description of the tools and processes followed to achieve the given results of the project report.

CHAPTER THREE

METHODOLOGY AND DESIGN

3.1 INTRODUCTION

This chapter demonstrates the entire methodology paradigm that would be used in this project work. The approach used in this project is completely data driven. In this project we applied several feature selection methods like information gain, gain ratio, relief F, and chi-squared to filter and reduce the data set and data set and again evaluate the accuracy of the prediction of the skin disease data set and again evaluate the accuracy of the prediction of the skin disease data set.

Then we applied classification algorithms such as support vector machine (SVM), Naïve Bayes, Decision Tree, k- nearest neighbor (kNN), Logistic Regression and Multi-Layer Perception (MLP) to measure the accuracy and sensitivity of the predicted values of skin disease classes.

The obtained values are than improved using several ensemble methods such as stacking (logical regression), bagging and boosting.

Here a combination method: voting scheme would be used to reconcile and get the best prediction.

3.2 DATASET

3.2.1 DATASET ANALYSIS

Data set is a collection of various types of data stored in a digital format. In any study analysis, exploratory data analysis (EDA) is a crucial stage. The main objective in order to direct specific testing for your hypothesis, exploratory analysis examines the data for distribution, outliers and anomalies.

3.2.2 DATA ACQUISITION

A single dataset was used for this analysis. The data set was obtained from the UCI machine learning repository (http://archive.ics.uci.edu/ml/dataset/dermatology), which is free and open to the public. The data set is to determine the type of Eryhemato- squamous disease.

The database contains 34 attributes, 33 of which are linear valued and one of them is nominal.

The dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicate the largest amount possible and 1, 2 indicate the relative intermediate values.

There are six classes of erythemato-squamous disease, with 366 instances and 34 features.

3.2.3 DATA PROCESSING

Before classification methods can be applied, the data must be cleaned and converted, which is referred to as the pre-processing stage. This pre-processing step involves, eliminating noisy data like outliers, normalizing and balancing imbalance data.

Depending on the kind of attribute, the mean or mode of the data set is substituted for the missing value using the mean imputation approach. They only discovered that various imputation strategies had varying effects on the classification accuracy methods.

The missing values in this project were filled in using the mean imputation approach. They only discovered that various imputation strategies had varying effects on the classification accuracy methods.

The missing values on this project were filled in using the mean imputation method. This choice as informed by the fact that 20% of the values in our data set were missing, as well as by how well it worked overall to improve the classification systems' accuracy.

A linear transformation of the data is applied using the Minimum-Maximum normalization technique. The minimum value that data can take is the smallest value, and the maximum value that data can accept is the largest value. The data often fall between 0 and 1.

In this project, minimum-maximum normalization was utilized to boost the algorithm's effectiveness and efficiency. Any variation in distance measurements where the data may not have been standardized will be avoided by this standardization (Olivera*et al*., 2017). The following formula in equation (3.1)

$$^{x}\text{scaled}=\frac{x-min\,(x)}{(x)-min\,(x)} \ldots\ldots\ldots\ldots(3.1)$$

where: x= single feature/variable

min (x)= minimum of x

max(x)= maximum of x

3.3 FEATURE SELECTED METHOD

In this project, after data processing, a hybrid feature selection method was used which comprises of four filter-based selection method; info gain, gain ratio, relief and chi-squared, and an embedded method; RFE-SVM to get the optimal dataset.

The following feature selection methods was used as stated below with their formulas:

3.3.1 INFORMATION GAIN

Information gain calculate the entropy decrease caused by a data set modification; by assessing each variables information gain in relation to the target variable, it can be used for feature selection.

The mean of individual trees' improvement in splitting criterion generated by each variable is essentially the feature importance.

The Gini impurity and entropy are the two most prevalent impurities. "Information gain" is an improvement on the entropy, whereas "Gini importance" is an improvement on the gini impurity (Verma *et al*., 2020). Gini imp (3.2) and entropy (3.3) can be calculated with equations:

Gini imp =

$$\sum_{i}\left(P_i(1-\ P_i)\right) ---- (3.2)$$

Entropy=

$$-\sum_{j} p_j\, log_2 p_j ----- (3.3)$$

3.3.2 GAIN RATIO

Information gain is modified by the gain ratio to lessen its bias. By considering the number of branches that will occur before splitting, gain ratio gets around the information gain issue. By accounting for the inherent information of a split, it corrects information gain.

Gain ratio is equal to the difference between the gaining partner's new and previous profit sharing ratios. Each gaining partner is computed independently (www.accountingcapital.com).

The formula for gain ratio is given in equation (3.4):

Gaining ratio = New profit-sharing ratio - Old profit sharing ratio……. (3.4)

3.3.3 RELIEFF

To the best of our knowledge, the relief algorithm and its derivatives are the only individual evaluation filter methods capable of spotting feature dependencies. These methods leverage the notion of nearest neighbors to obtain feature statistics that cover interaction without explicitly searching through feature combinations.

Relief F randomly choses an instance R_1 from class. It can find K for the nearest neighbors from the same class (nearest hits H) and from the different classes (nearest misses M) (Michalska., 2021).

i.e.

$$W_i = W_i \frac{\sum^{k} D_H(k)}{n \cdot k} + \sum_{c=1}^{c=1} P_c \cdot \frac{\sum_{k=1}^{k} D_m(k)}{n \cdot k} \quad \ldots\ldots (3.5)$$

3.3.4 CHI-SQUARE

One of the primary feature selection processes is the chi-square feature selection method is employed to determine the relationship between qualities and the target attribute. Statistics uses the chi-square test to examine the consistency of two events. In feature selection, in particular, we utilize chi-square to determine whether there is an independent relationship between the presence of a particular attribute and the occurrence of a particular class (Verma *et al*., 2020).

Chi-square values are calculated using formula (3.6)

$$x_c^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} \ldots\ldots\ldots (3.6)$$

Where, c = degrees of freedom

O = observed value

E = expected value

3.3.5 RECURSIVE FEATURE ELIMINATION SUPPORT VECTOR MACHINE (RFE-SVM)

The RFE-SVM works by searching for a subset of a feature in the training dataset and successfully removing features until the desired number remains (Rustam and Kharis., 2020). Equation (3.7):

$$w = \sum_{k \epsilon SV} Y_k \alpha_k X_k \ldots\ldots\ldots (3.7)$$

Where w = weight of vector

Y_k= corresponding to target, $Y_k = \{\pm 1\}, i = 1, \ldots, m$

X_k = Lagrange coefficients

X_k= training data $X_k \in R^n$

3.4 CLASSIFIFICATION ALGORITHMS

After the optical data set is obtained via the hybrid feature selection process, then classification algorithms are used to process the sub-data set. The following classification was used in this proposed model:

3.4.1 SUPPORT VECTOR MACHINE

Support vector machine are supervised learning algorithms that examines the data used classification and regression analysis as well as outlier detection. Discriminant classifiers called

support vector machines are accurately described by distinct hyper planes, the SVN classifiers, the new recognition using labeled training data (supervised mama) and the optical hyper plane are the support vectors (Verma and Pal., 2019). Equation (3.8) below shows the formula for SVM.

$$x.y = \sum_{i=1}^{h} x_i y_1 \text{------} (3.8)$$

3.4.2 DECISION TREES (DTs)

An example of supervised machine learning is decision trees. An algorithmic method for dividing a data set into segments based on several criteria is used to build decision trees. The objectives is to learn straightforward decision rules derived from the data features in order to build a model that predicts the value of a target variable (Verma *et al*., 2019).

It creates an associated decision tree by breaking up a dataset into numerous smaller data subsets. As a result, a tree with decision nodes which as two or more branches and leaf nodes that reflect classifications and decisions is produced (Verma *et al*., 2019).

How the Decision Tree Algorithm Works:

The choice to make strategic splits has a significant impact on a tree's accuracy. Regression and classification tress have different decision criteria.

To decide whether to divide a node into two or more sub-decisions, trees employ a variety of techniques. The homogeneity of newly formed sub nodes is increased by sub-node creation. In other words, we can claim that the nodes purity improves in relation to the desired variables that are accessible before choosing the split that produces the most homogenous sub-nodes.

Entropy is a metric used in information theory to gauge how pure or uncertain a set of observation is. It controls the decision tree's choice on how to divide the data (www.Kdnuggets.com).

$$E(S) = \sum_{I=1}^{C} - P_i log_2 P_i \ldots\ldots\ldots\ldots\ldots (3.9)$$

3.4.3 k-NEAREST NEIGHBOUR (KNN)

Although kNN is typically employed in classification problems, kNN base learner is also utilized for regression challenges. In situations where the datasets don not match hypothetical mathematical assumptions, kNN will be useful for prediction (Verma *et al*., 2020).

It is a supervised learning technique as well. With this technique, each sample was grouped with nearby samples that were comparable to it. As a result, the samples in its immediate surroundings can be used to classify an unknown sample. The classifier determines the distance between the unknown sample and all datasets when a training dataset and an unknown sample are provided. The training set and the unknown sample are separated by the value with the least value being accepted.

When classifying an unknown sample, the effectiveness of kNN mostly depends on the value of K, the number of nearest neighbours. The categorization is not very good if k is quite small. However, the greatest number of k yields the ideal classification result (Bose *et al*., 2021).

Fundamentally, the k – nearest neighbor classifier depends on a distance metric, the more accurately that metric captures label similarly, the more accurate the classification. The minkowski distance is the most popular option (www.cs.cornell.edu).

$$dist(x,z) = \left(\sum_{r=1}^{d} |x_r - z_r|^P \right)^{\frac{1}{p}} \text{.....(3.10)}$$

3.4.4 MULTILAYER PERCEPTION (MLP)

A logical regression classifier called a multilayer perception (MLP) is used. This classifier alters input data via a learned nonlinear conversion. This transforms the incoming data into a layer that is linearly divisible. A hidden layer is used, otherwise, it functions as an ANN (Artificial Neural Network). Although there are several advantages to using numerous hidden layers. (Verma *et al*., 2020).

Because MLPs can solve issues stochastically, which frequently enables approximate solutions for incredibly difficult problems like fitness approximation, they are useful in research.

Each layer is represented as:

$$y = f(WxT + h) \text{ ….. } (3.11)$$

Where x is the input vector, which can also be the output of the preceding layer, f is the activation function, W is the set of parameter or weights in the layers and b is the bias vector (www.en.m.wikipedia.org).

3.4.5 NAIVES BAYES

The number of parameters required for naïve Bayes classifiers is linear in the number of variables (feature/predictors) in a learning problem, making them extremely scalable. Instead of using an expensive iterative approximation, as is the case for many other types of classifiers maximum-likelihood training can be performed in linear time by evaluating a close form expression, which is faster.

Naïve Bayes is a straightforward method for building classifiers. These model assign class labels to problem cases, which are represented as vectors of feature values and the class labels are chosen from a finite set.

A problem instance to be classified is represented by a vector $x = (x_1, ..., x_n)$ encoding some n features (independent variables); Naïve Bayes is an abstract conditional probability model that assigns probabilities. $P((\underset{x}{K}—, ...x_n)$ to this issue instance for each of k potential outcomes or classes (www.en.m.Wikipedia.org). Using Bayes theorem, the condition probability can be decomposed as;

$$p((k|x) = \frac{p((k)p(x|(k)}{p(x)} \text{ …… } (3.12)$$

3.4.6 LOGISTIC REGRESSION (Meta –Classifier)

The logistic model, often known as the logit model, is a statistical model that estimates the likelihood of an event occurring by making the events log-odds a linear combination of one or more independent variables.

In regression analysis, logistics regression (also known as logit regression) estimates a logistic model's parameters (the coefficients in the linear combination). Formally, binary logistic regression has a single binary dependent variable (two classes, coded by an indicator variable) with the value "0" and "1", while the independent variables can either be continuous variables or binary variables (two classes, coded by an indicator variable) (any real value)

Logistic regression is generally use in different fields. Including machine learning, most medical fields, and social sciences.

The logical functions is of the form:

$$P(x) = \frac{1}{1+e^{-(x-\mu)/5}} \ldots\ldots (3.13)$$

Where μ is a location parameter (the midpoint of the curve, where p(μ) = ½ and s is a scale parameter.

Logistic regression is used as the meta-classifier which makes the final prediction among all the predictions of other classifiers mentioned earlier by using those predictions as features (www.en.m.wikipedia.org).

3.5 ENSEMBLE LEARNING TECHNIQUE

Several base learners are combined using ensemble methods to predict a problem, improving the prediction of a single classifier. The two main classifiers of ensemble methods are joining multiple classifiers of different types.

3.5.1 STACKING

Combining many base learners of various types with the aid of a meta-classifier is done using the stacking ensemble method.

A fresh data set from the first-level classifiers must be produced for the stacking training phase. There is substantial risk of over fitting of the exact same data used to train the first-level learners are also utilized to generate the fresh data set for training the second-level trainers. A cross-validation or leave-one-out technique is frequently advised, and it is proposed that the instances

used to create the new data set be excluded from the training examples for the first-level trainers (Zhou, Z.-H. 2012).

The pseudo-code of a general stacking procedure is summarized in figure 3.1.

$$Input: Data\ set\ D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\};$$
$$First - level\ algorithms\ \mathcal{L}_1, \ldots\ldots\ldots, \mathcal{L}_T;$$
$$Second - level\ learning\ algotithm\ \mathcal{L}.$$

$Process$;
1. $For\ t = 1, \ldots, T$: % Train a first-level learner by applying the
2. $ht = \mathcal{L}_t(D)$: % First-Level learning algorithm $\mathcal{L}_t$
3. end
4. $D' = \emptyset$: % Generate a new dataset
5. $For\ i = 1, \ldots, m$:
6. $For\ t = 1, \ldots, T$:
7. $z_{it} = h_t(x_i)$:
8. end
9. $D' = D'V\big((z_{i1}, \ldots, z_{iT})y_i\big)$;
10. end
11. $h' = \mathcal{L}\big(D'\big)$; % Train the second-level learner h' by
% applying the second-level learning
% algorithm $\mathcal{L}$ to the new dataset D'.

$Output: H(x) = h'(h_1(x), \ldots, h_{T(x)})$

Fig 3.1: A general stacking procedure

As an illustration, consider K-fold cross- validation. In this case the original training data set D is arbitrarily divided into K almost equal parts D1, ----,$D_{k.}$ define D_j and D(-j) =D/Dj to get the test and braining sets for the jth fold . by calling the tth learning algorithms, on D(j) with respect to learning algorithms, a first learner h(j) is produced. let z stand for the learner h(j) on x output for each xi in Dj, the test set of the jth fold. The new data set is generated from the T individual learners as:

$$D' = \{(z_{i1}, \ldots, z_{iT}, Y_i)\}_{i=1}^{m}, \ldots\ldots \quad (3.14)$$

The learner h' produced by applying the second level learning algorithm is a function of (Z1,--, ZT) for Y. the final first level learners are often generated again by training on the entire training data after creating the new data set (Zhou, Z.-H. 2012).

3.5.2 BAGGING

The two key ingredients of bagging are bootstrap and aggregation. Bagging is an ensemble method that splits the training set into groups and creates a classifier for each group. Each group of the training dataset is different. It takes the average or the majority voting to combine the multiple classifier results (Zhou, Z.-H. 2012). The bagging algorithm is summarized in fig 3.2

$$Input: Data\ set\ D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\};$$
$$Base\ learning\ algorithm\ \mathcal{L};$$
$$Number\ of\ base\ learners\ T.$$

$Process;$
1. $For\ t = 1, \ldots, T:$
2. $ht = \mathcal{L}(D, D_{bs})$: % Dbs is the bootstrap distribution
3. end

$$Output: H(x) = \underset{y \in Y}{\text{argmax}} \sum_{t=1}^{T} \mathbb{I}(h_t(x) = Y)$$

Fig 3.2 The Bagging Algorithm

We must keep track of training examples utilized for each base learners in order to obtain the out-of-bag estimate. The out-of –bag prediction on x is devoted as $H^{oob}(x)$, and only learners who have not been trained on x are involved, i.e:

$$H^{oob}(x) = \underset{y \in Y}{\text{argmax}} \sum_{t=1}^{T} \mathbb{I}(h_t(x) = Y) \cdot \mathbb{I}\ (x \notin D_t) \ldots. (3.15)$$

Then, the out-of-bag estimate of the generalization error of bagging is:

$$err^{oob} = \frac{1}{|D|} \sum_{(x,y) \in D} \mathbb{I}\left(H^{oob}(x) \neq Y\right) \ldots.. (3.16)$$

3.5.3 BOOSTING

A class of algorithms known as boosting are capable of transforming weak learners into strong learners. A strong learner is extremely close to flawless performance, while a weak learner is only marginally better than a random guess.

Briefly stated boosting involves teaching a group of learners in a sequential manner and combining them for prediction, with the later learners concentrating more on the mistakes of the earlier learners (Zhou, Z.-H. 2012). The general boosting procedures is summarized in fig 3.3

$Input: Sample\ distribution\ D;$
$Base\ learning\ algorithm\ \mathcal{L};$
$Number\ of\ learning\ the\ rounds$ T.

$Process;$
1. $D_1 = D.$ % Initialize distribution
2. $for\ t = 1, \ldots, T:$
3. $h_t = \mathcal{L}(D_t);$ % Train a weak learner for distribution, D_t
4. $\in_t = P_x \sim D_t\big(h_t(x_1) \neq f(x)\big);$ % Evaluate the error of h_t
5. $D_{t+1} = Adjust\_Distribution\ (D_t, \in_t)$
6. end

$Output: H(x) = Combine\ outpurs\ (\{h_1(x), \ldots., h_{t(x)\}})$

Fig 3.3: A general boosting procedure

3.5.3.1 ADABOOST ALGORITHM

Since it include component like adjust distribution and combine outputs that are not explicitly stated, the general boosting procedure shown in fig 3.1 is not a true algorithm. The most famous boosting algorithm, the adaboost algorithm can be seen as an instantiation of these components in fig 3.4.

$Input$: $Sample\ distribution\ D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
$Base\ learning\ algorithm\ \mathcal{L}$;
$Number\ of\ learning\ the\ rounds$ T.

$Process$;

1. $D_1(x) = \frac{1}{m}$. % Initialize the weight distribution
2. $For\ t = 1, \ldots, T$:
3. $h_t = \mathcal{L}(D, D_t)$; % Train a classifier h_t from D under distribution D_t
4. $\in_t = P_x \sim D_t\big(h_t(x_1) \neq f(x)\big)$; % Evaluate the error of h_t
5. $If \in_t = 0.5, then\ break$
6. $\alpha_t = \frac{1}{2}\ In\ (\frac{1-\in_t}{\in_t})$ % Determine the weight of h_t
7. $D_{t+1}(x) = \frac{D_t(x)}{z_t} \times \begin{cases} \exp(-\alpha t)\ if\ h_t(x) = f(x) \\ \exp(\alpha t)\ if\ h_t(x) \neq f(x) \end{cases}$
$= \frac{D_t(x) \exp(-\alpha t f(x)\, h_t(x))}{z_t}$ %Update the distribution
% where $z_t$is a normalization
% factor which enables D_{t+1} to
% be a distribution.
8. end

$Output$: $H(x) = sign\ (Z_t^T = 1\ \alpha_t h_t(x))$

Fig 3.4: The AdaBoostSytem

The solution is:

$$\alpha_t = \frac{1}{2} In \left(\frac{1-\epsilon_t}{\epsilon_t}\right) \qquad - \qquad - \qquad (3.12)$$

3.6 COMBINATION METHOD

Ensemble method use combination to produce a good generalization ability after generating a series of base learners rather than attempting to discover the best single learner, where the combination approach plays a significant role (Zhou, Z.-H. 2012).

3.6.1 VOTING SCHEME

A voting ensemble (or a" majority voting ensemble") is an ensemble machine learning model that combines the predictions from multiple pother models.

It is a technique that may be used to improve model performance, ideally achieving better performance than any single model used in the ensemble.

There are two approaches to the two majority vote prediction for classification; they are

1. Hard voting: predicts the class with the largest form of votes from model

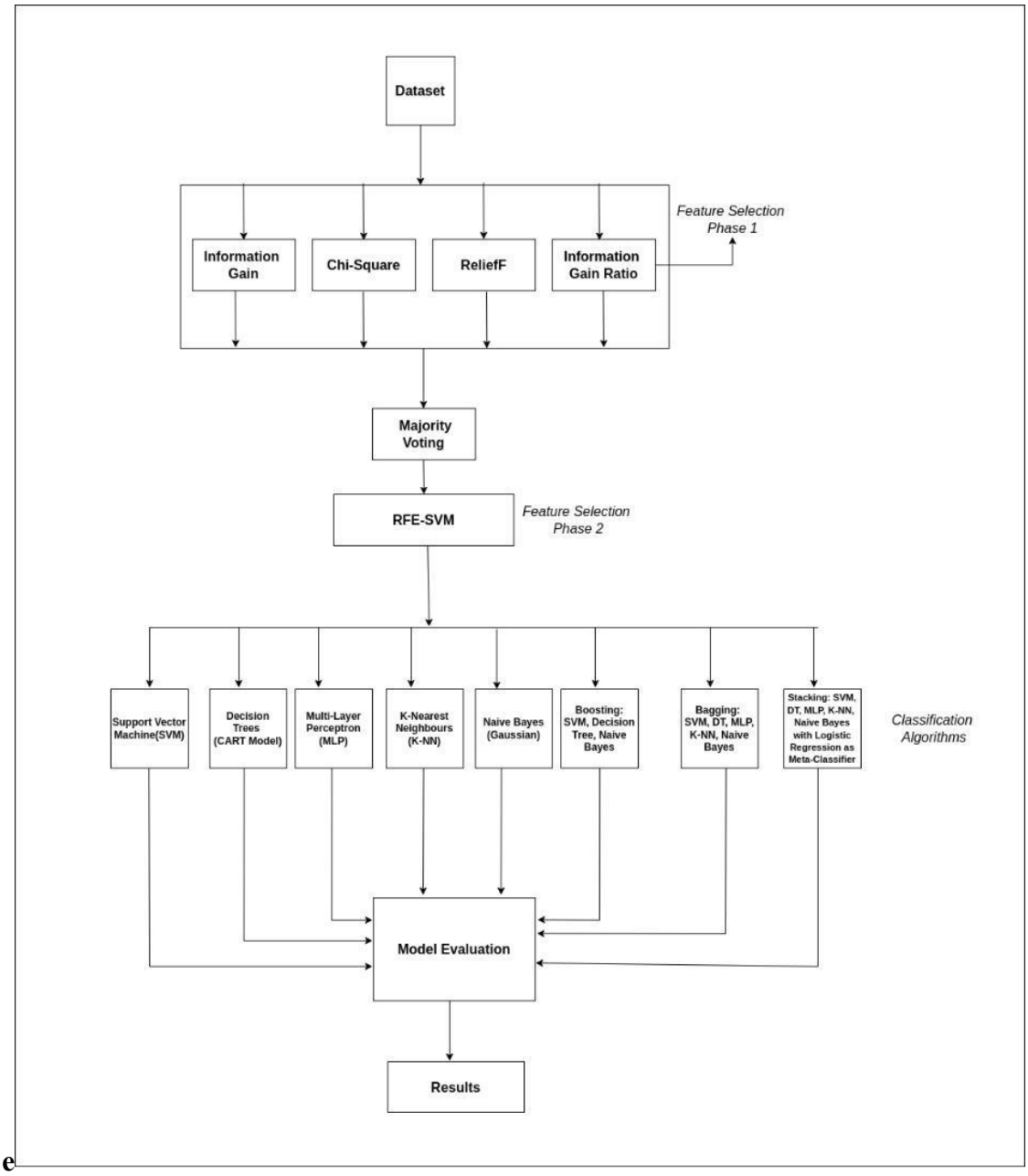
2. Soft voting: predict the class with the largest summed probability from models.

This project makes use of hard voting techniques as its majority voting. Majority voting is the most popular voting method. Here, every classifier votes for one class label, and the final output class label is the one that receives more than half of the votes, a rejection option will be given and the combined classifier makes no prediction (Zhou, Z.-H. 2012). That is, the output class label of the ensemble is:

$$H(x) = \begin{cases} C_j \sum_{i=1}^{T} h_i^j(x) > \frac{1}{2}\sum_{i=1}^{T} h_i^k(x) \ldots\ldots (3.18) \\ \text{rejection} \qquad \text{otherwise} \end{cases}$$

Where T is the set f individual classifiers, $\{h_1, \ldots h_i, \}$ is the possible class labels, $\{C_1, \ldots C_i, \}$ are the instances, h_i are the output of the classifiers and h_i^i is the output of h_i for the class labels C_i(Zhou, Z.-H. 2012).

3.7 PROPOSED DESIGN



e

Fig 3.1: Architecture of Proposed System

3.8 PERFORMANCE EVALUATION

In order to determine the effectiveness of the classification algorithm used, a measurement is needed. Commonly used measurements include classification accuracy, F-Measure, precision, recall, Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC). These measurements can be calculated by the classification results commonly tabulated in a matrix format called a Confusion Matrix. In a classic binary classification problem, the classifier labels the items as either positive or negative (Villacampa, 2015). A confusion matrix summarizes the outcome of the algorithm in a matrix format (Chawla, 2005). In our binary example, the confusion matrix would have four outcomes:

a) True positives (TP) are positive items correctly classified as positive.

b) True negatives (TN) are negative items correctly identified as negatives.

c) False positives (FP) are negative items classified as positive.

d) False negatives (FN) are positive items classified as negative.

The confusion matrix is represented in Table 3.1.

Table 3.1. Confusion matrix

Confusion matrix		**Classified As:**	
		Negative	**Positive**
Actual Class	**Negative**	**TN**	**FP**
	Positive	**FN**	**TP**

The following performance measures use the values of the confusion matrix in their calculation.

1. Classification Accuracy:

The simplest performance measure is accuracy. The overall effectiveness of the algorithm is calculated by dividing the correct labeling against all classifications.

$$Acc = \frac{TN+TP}{TN+TP+FN+FP} \quad \ldots \quad (3.19)$$

The accuracy determined may not be an adequate performance measure when the number of negative cases is much greater than the number of positive cases (Kubat and Matwin, 1997)

2. F-Measure

F-Measure is one of the popular metrics used as a performance measure. The measure itself is computed using two other performance measures, precision and recall.

$$Precision = \frac{TP}{TP+FP} \quad \ldots \quad (3.20)$$

$$Recall = sensitivity = \frac{TP}{TP+FN} \quad \ldots \quad (3.21)$$

Precision is the number of positive examples classified over all the examples classified. Recall, also called the True Positive Rate (TPR), is the ratio of the number of positive.

Precision is the number of positive examples classified over all the examples classified. Recall, also called the True Positive Rate (TPR), is the ratio of the number of positive examples classified over all the positive examples. Based on these definitions F-measure is defined as follows:

$$f - measure = \frac{2*precision*recall}{Precision + recall} \quad \ldots \quad (3.22)$$

In essence, the F-Measure is the harmonic mean of the recall and precision measures.

3. Sensitivity and Specificity

The performance of a binary classifier may sometimes be quantified by its accuracy as

described above, i.e. the portion of misclassified classes in the entire set. However, there

may be times when the types of misclassifications may be crucial in the classification assignment (Powers, 2011). In these cases, the values for sensitivity and specificity are used in determining the performance of the classifier. Sensitivity or Recall or True

Positive Rate (TPR) is the ratio of true positive predictions over the number of positive instances in the entire data set.

$$Sensitivity = \frac{TP}{TP+FN} \quad \ldots \quad (3.23)$$

The specificity or True Negative Rate (TNR) is the ratio of true negative predictions over the number of negative instances in the entire data set.

$$Specificity = \frac{TN}{TN+FP} \quad \ldots \quad (3.24)$$

These values can be further analyzed using a Receiver Operating Characteristic Curve (ROC) where the sensitivity is plotted against 1- specificity (Fawcett, 2006). ROC is described further in the next section.

CHAPTER FOUR

IMPLEMENTATION AND DOCUMENTATION

4.1 OVERVIEW

The actual creation and testing of the model is referred to as system implementation. The purpose of this part is to build the suggested model, document the process, and provide the results.

4.2 SYSTEM REQUIREMENTS

System requirements are the settings that a system must have in order to function smoothly, effectively, efficiently, and predictably, with the understanding that failing to fulfil these requirements might have an impact on performance or cause issues.

4.2.1 HARDWARE REQUIREMENTS

A good laptop with the following hardware is needed to run the model conveniently:

a. Processor: Intel 13, 15, 17, and above
b. RAM: 12GB or more
c. Memory: above 120GB of space
d. Processor Speed: 2.2GHZ and above
e. GPU: Because the model uses several cores for processing during its training stages, a GPU is necessary to increase the model's performance.
f. TPU: Because the model uses several cores for processing during its training stages, a TPU is necessary to increase the model's performance.
g. Internet: A stable internet connection from a reliable Internet Service Provider (ISP) because the software used 1s required to run online.

4.2.2 SOFTWARE REQUIREMENTS

A computer's software is an intangible Component of the computer. To run the model, the following software is required:

a. Operating System: either a Windows 7, 8, 10, or Linux operating system should be sufficient but Linux is recommended,

b. Integrated Development Environment (IDE): Google Colaboratory (Google Colab)

4.3 TOOLS FOR MODEL DEVELOPMENT

Software developers use these programs to create, debug, and maintain other programs and applications. The next section lists the most important software tools utilized in the creation of the proposed model.

4.3.1 PROGRAMMING LANGUAGES USED

In the development of a model, the decision of which programming language to use is critical since it helps the programmer to convey his or her ideas in a convenient manner.

A. Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a Scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form. The proposed model utilizes some of python’s extensive library which includes:

1. Numpy is a python library that allows provides its users with an array processing package. It allows for the use of high-performance multidimensional array object.
2. Pandas is used as a data manipulation tool that makes use of a data frame in storing its data in a tabular form.
3. Seaborn is used for data visualization. It is a high-level interface based on matplotlib (another popular data Visualization library). It can be used in all sorts of data analysis tasks that require visualization of data and inferring information from it.
4. Matplotlib is a library used in visualizing data by providing the ability to create graph plots

B. Scikit-Learn

Another powerful library used in the proposed model is Scikit-Learn. Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. It is built upon some of the familiar technologies like NumPy, pandas, and Matplotlib. The functionality that scikit-learn provides includes:

a. Regression, including Linear and Logistic Regression

b. Classification, including K-Nearest Neighbors

C. Clustering, including K-Means and K-Means++

d. Model selection

e. Preprocessing, including Min-Max Normalization

4.4 SYSTEM TESTING

System testing is used to assess the model's performance. In general, model performance testing comprises running the models against the test dataset and evaluating the outcomes in terms of accuracy, recall, and precision.

4.5 RESULTS AND DISCUSSION

This section includes images of some of the model's intriguing feature graphs, as well as the suggested model's outcome at the end.

Raw Dataset: The data set was obtained from the data repository at the University of California, Irvine (UCI). The raw dataset is shown in figure 4.1 below. The dataset shows the different data characteristics.

Processed Dataset: The data set was obtained from the data repository at the University of California, Irvine (UCI). The processed dataset shown in figure 4.2 shown below depicts the dataset that is processed using equation 1.1 in normalizing the data for machine

learning. The essence of preprocessing in the form of normalizing the dataset is to reduce variance in the dataset that is to use a common scale without distorting differences in the ranges of values.

Table 4.1: The raw

Erythema	scaling'	definite borders'	itching'	phenomenon'	polygonal papules'	follicular papules'	mucosal involvement'	elbow involvement'	scalp involvement'	family history'	incontinence'	in the infiltrate'	PNL infiltrate'	the papillary dermis'	exocytosis'	acanthosis'	hyperkeratosis'	parakeratosis'	the rete ridges'	of the rete ridges'	suprapapillary epidermis'	spongiform pustule'	microabcess'	hypergranulosis'	granular layer'	damage of basal layer'	spongiosis'	appearance of retes'	follicular horn plug'	parakeratosis'	monocuclear infiltrate'	band-like infiltrate'	Age'
2	2	0	3	0	0	0	0	1	0	0	0	0	0	0	3	2	0	0	0	0	0	0	0	0	0	3	0	0	0	1	0	55
3	3	3	2	1	0	0	0	1	1	1	0	0	1	0	1	2	0	2	2	2	2	2	1	0	0	0	0	0	0	0	1	0	8
2	1	2	3	1	3	0	3	0	0	0	1	0	0	0	1	2	0	2	0	0	0	0	0	2	0	2	3	2	0	0	2	3	26
2	2	2	0	0	0	0	0	3	2	0	0	0	3	0	0	2	0	3	2	2	2	2	0	0	3	0	0	0	0	0	3	0	40
2	3	2	2	2	2	0	2	0	0	0	1	0	0	0	1	2	0	0	0	0	0	0	0	2	2	3	2	3	0	0	2	3	45
2	3	2	0	0	0	0	0	0	0	0	0	2	1	0	2	2	0	2	0	0	0	1	0	0	0	0	2	0	0	0	1	0	41
2	1	0	2	0	0	0	0	0	0	0	0	0	0	3	1	3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	18
2	2	3	3	3	3	0	2	0	0	0	2	0	0	0	2	3	0	0	0	0	0	0	0	0	2	2	3	2	0	0	3	3	57
2	2	1	0	2	0	0	0	0	0	0	0	0	0	0	2	1	0	1	0	0	0	0	0	0	0	0	2	0	0	0	2	0	22
2	2	1	0	1	0	0	0	0	0	0	0	0	0	0	3	2	0	2	0	0	0	0	0	0	0	0	2	0	0	0	2	0	30
3	3	2	1	1	0	0	0	2	2	1	0	0	0	0	0	3	2	3	2	2	2	1	1	0	0	0	0	0	0	0	1	0	20
2	2	0	3	0	0	0	0	0	0	0	0	0	2	0	2	2	0	0	0	0	0	1	0	0	0	0	3	0	0	0	1	0	21
3	3	1	2	0	0	0	0	0	1	0	0	0	2	0	3	1	0	1	0	0	0	0	0	0	0	0	2	0	0	0	1	0	22
2	3	3	0	0	0	0	0	1	1	1	0	0	1	0	0	2	1	2	1	2	3	0	2	0	0	0	0	0	0	0	2	0	10
2	2	3	3	0	3	0	2	0	0	0	2	0	0	0	1	1	1	1	0	0	0	0	0	2	0	3	0	3	0	0	1	3	65

	Erythema	scaling	definite borders	itching	koebner phenomenon	polygonal papules	follicular papules	oral mucosal involvement	knee and elbow involvement	scalp involvement	disappearance of the granular layer	vacuolisation and damage of basal layer	spongiosis	saw-tooth appearance of retes	follicular horn plug	perifollicular parakeratosis	inflammatory monoluclear inflitrate	band-like infiltrate	Age	Class
0	2	2	0	3	0	0	0	0	1	0	0	0	3	0	0	0	1	0	55	2
1	3	3	3	2	1	0	0	0	1	1	0	0	0	0	0	0	1	0	8	1
2	2	1	2	3	1	3	0	3	0	0	0	2	3	2	0	0	2	3	26	3
3	2	2	2	0	0	0	0	0	3	2	3	0	0	0	0	0	3	0	40	1
4	2	3	2	2	2	2	0	2	0	0	2	3	2	3	0	0	2	3	45	3

Features Selected by the Phase 1 Feature Selection Algorithms:

CHI-SQUARE: Figure 4.1 illustrate the diagram for chi-square feature selection method. The original 34 features are captured in figure 4.1 but only 22 features from chi-squared. The heuristic method is used to determine the number of features, i.e. human user defined method based on their rankings.

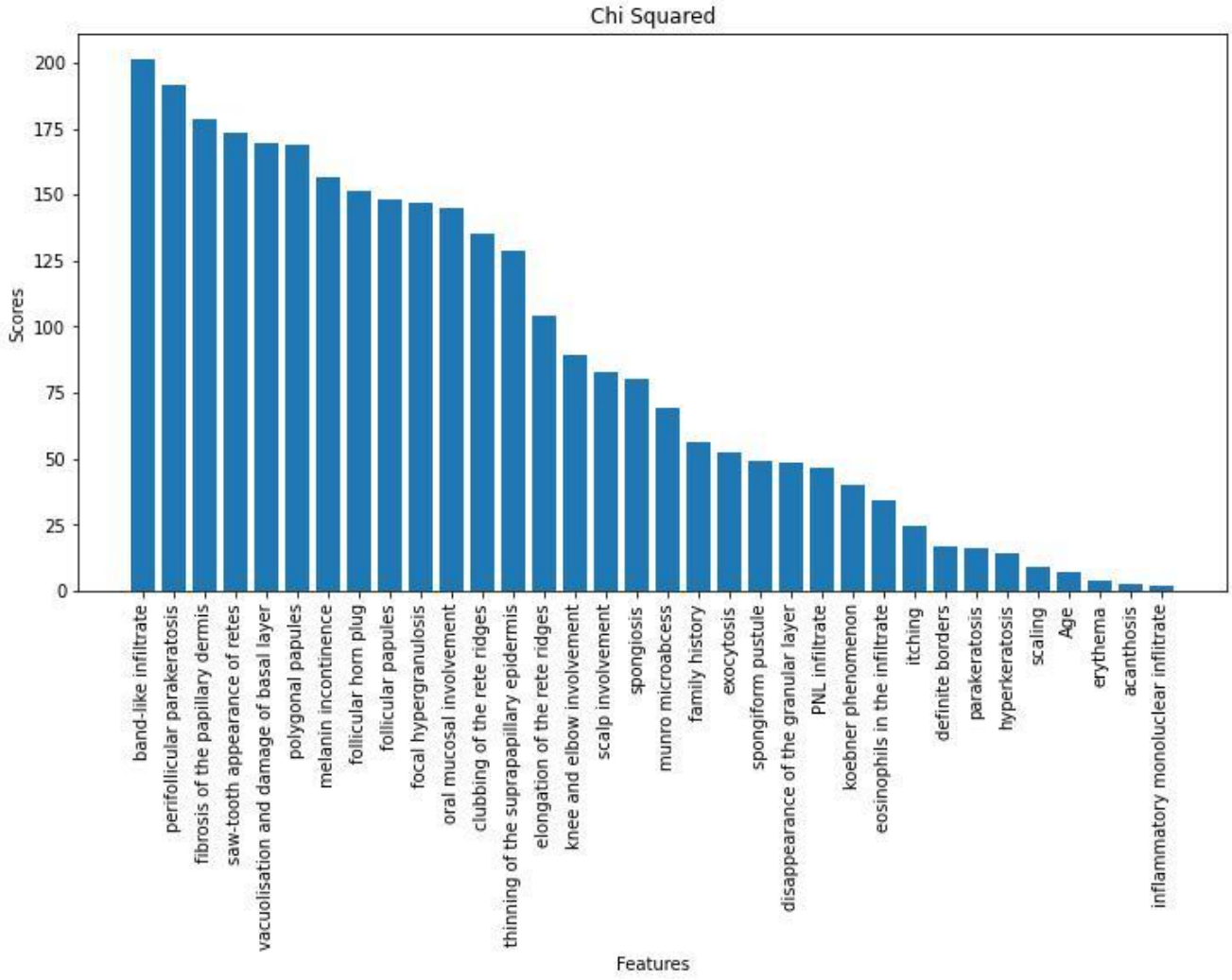


Fig 4.1. Chi-Squared Feature selection method

INFORMATION GAIN: Figure 4.2 illustrate the diagram for information gain feature selection method. The original 34 features are captured in figure 4.2 but only 22 features from information gain. The heuristic method is used to determine the number of features, i.e. human user defined method based on their rankings.

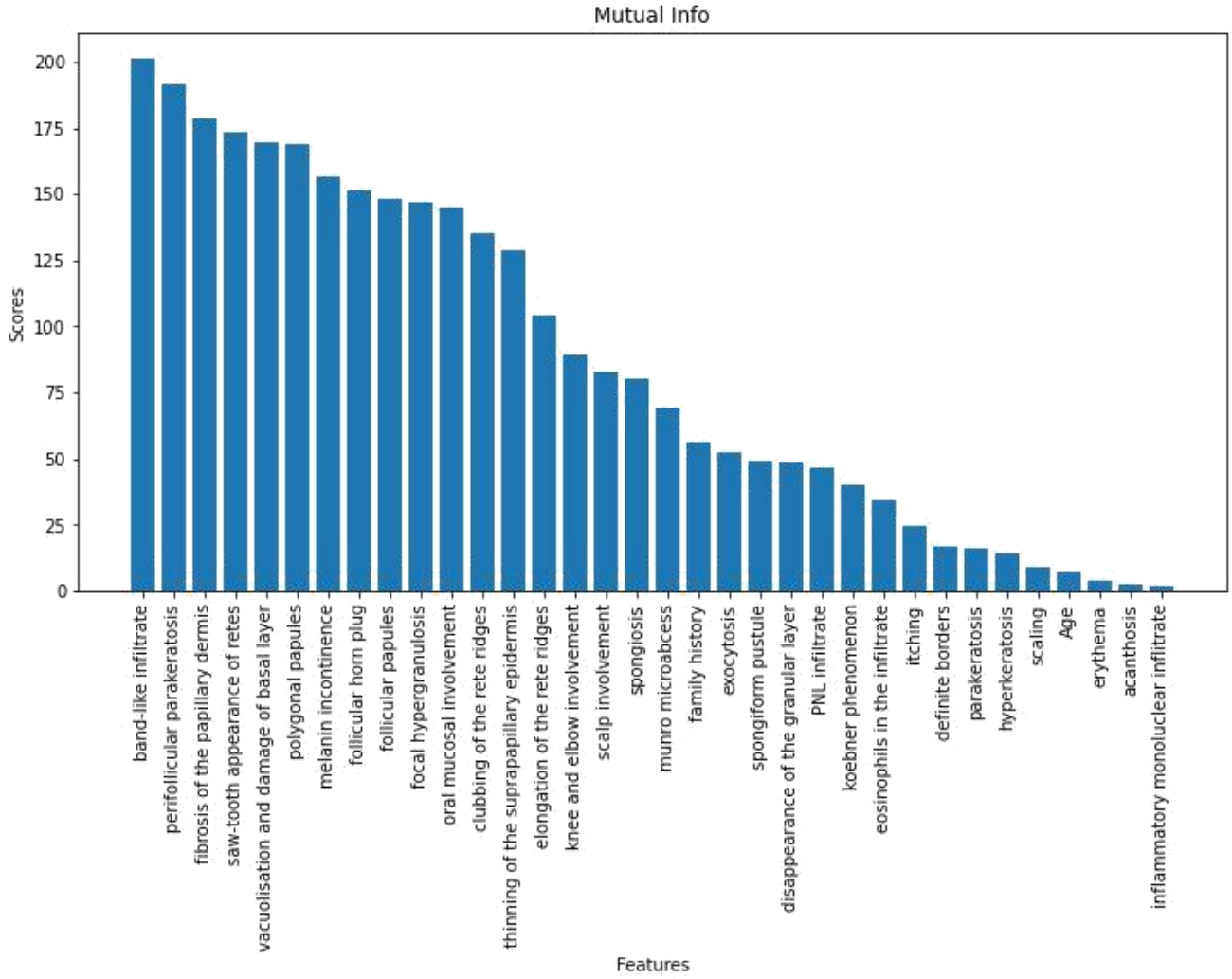


Fig 4.2. Information Gain Feature selection method

GAIN RATIO: Figure 4.3 illustrate the diagram for gain ratio feature selection method. The original 34 features are captured in figure 4.3 but only 22 features from gain ratio. The heuristic method is used to determine the number of features, i.e. human user defined method based on their rankings.

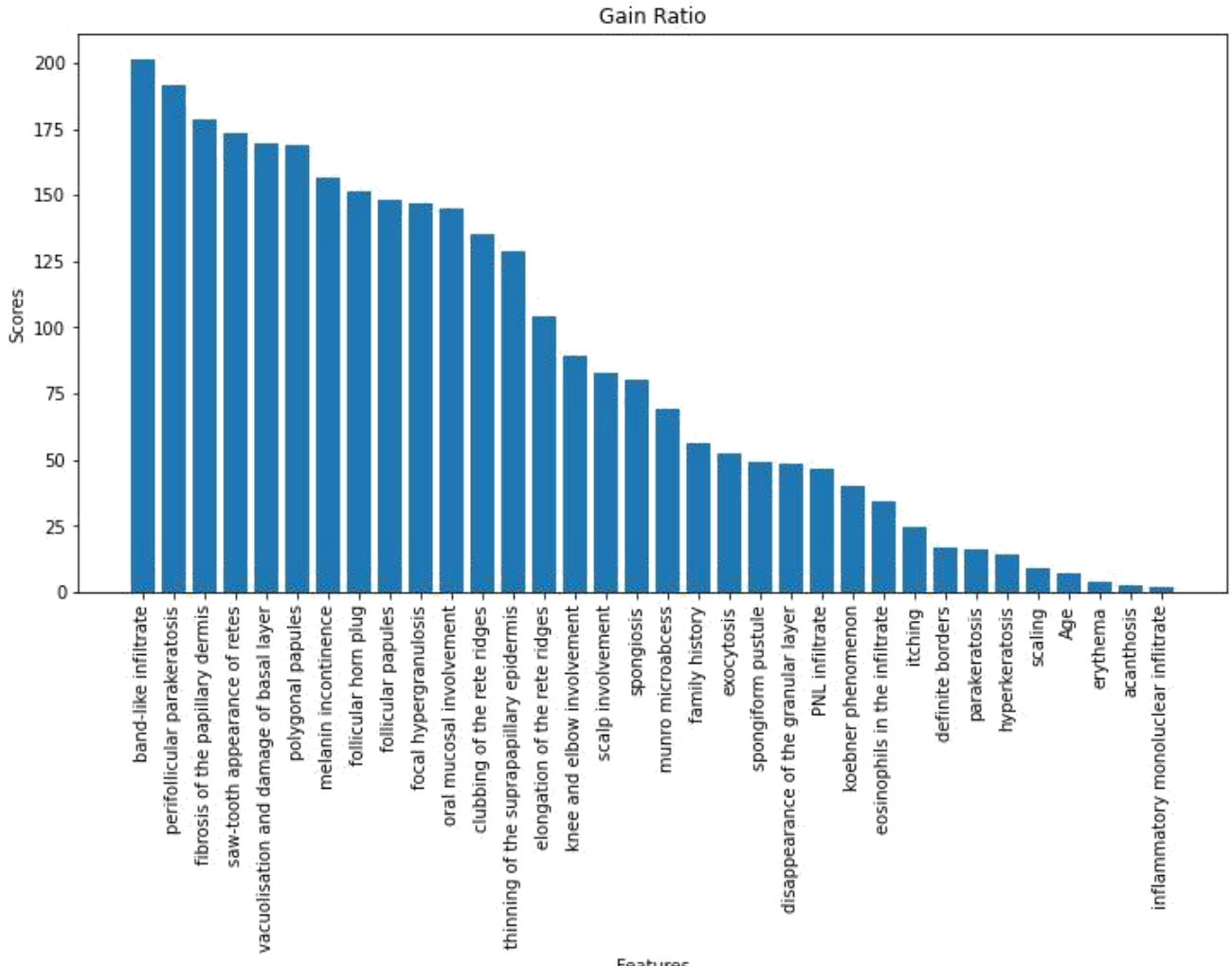


Fig 4.3. Gain Ratio Feature selection method

RELIEFF: Figure 4.4 illustrate the diagram for reliefF feature selection method. The original 34 features are captured in figure 4.4 but only 22 features from reliefF. The heuristic method is used to determine the number of features, i.e. human user defined method based on their rankings.

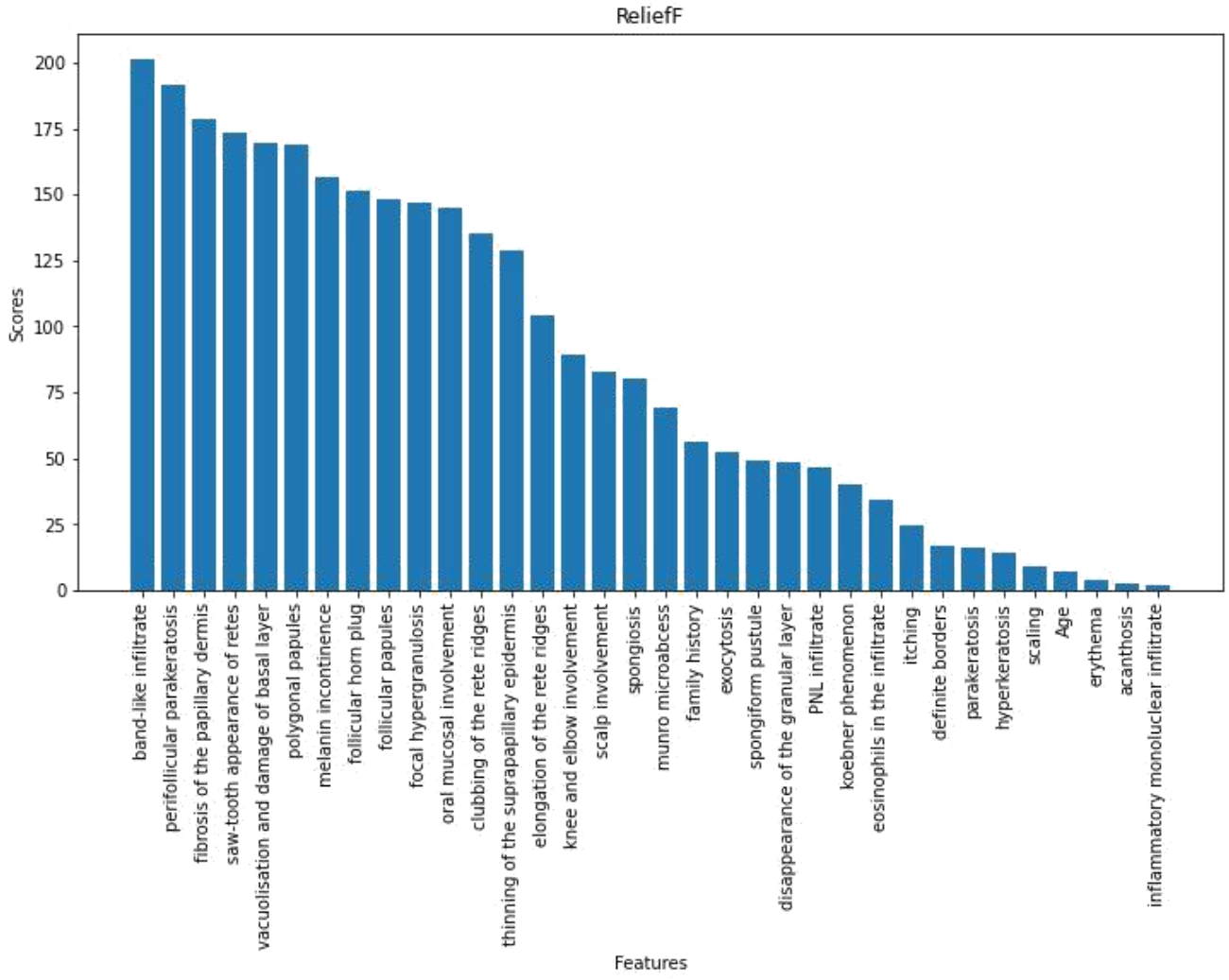


Fig 4.4. ReliefF Feature selection method

ENSEMBLE OF FILTER-BASED FEATURE SELECTION USING RFE-SVM:

Table 4.3 illustrates the final feature subsets from the combination of the filter methods i.e. chi-squared, information gain, gain ratio, and relief feature selection methods. The RFE-SVM is used to select 12 final feature subsets using recursive feature elimination method through SVM weightings.

Table 4.3: 12 Final Feature Subsets

NO OF FEATURE SUBSETS	FINAL FEATURE SUBSET SELECTED
1	Band-like infiltrate
2	Perifollicular parakeratosis
3	Fibrosis of the papillary dermis
4	Polygonal papules
5	Follicular papules
6	Focal hypergranulosis
7	Clubbing of the rete ridges
8	Thinning of the suprapapillary epidermis
9	Elongation of the rete ridges
10	Spongiosis
11	Exocytosis
12	Disappearance of the granular layer

CONFUSION MATRIX: Figure 4.5 to 4.18 shows the different confusion matrices for the various models used in this project. The confusion matrix is used as a technique to summarize the different classification algorithms performance.

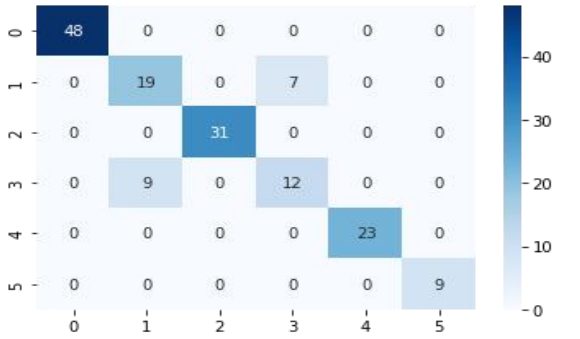


Fig 4.5: SVM

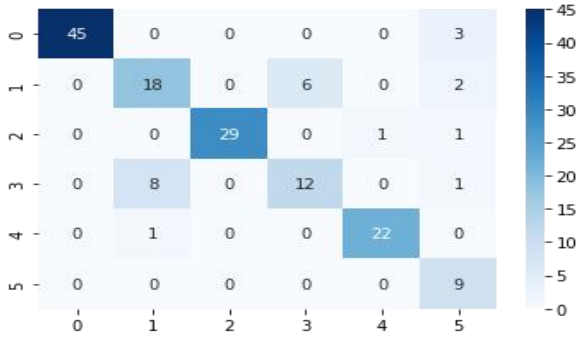


Fig 4.6: Decision Tree

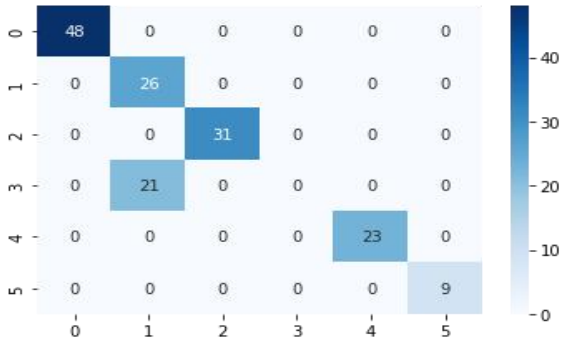


Fig 4.7: MLP

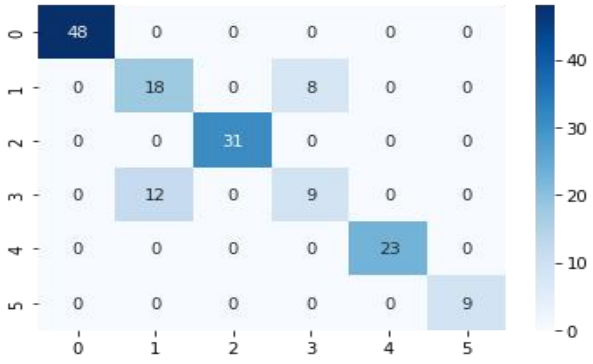


Fig 4.8: KNN

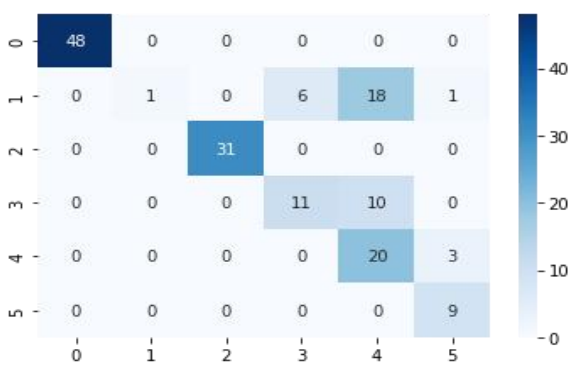


Fig 4.9: Gaussian NB

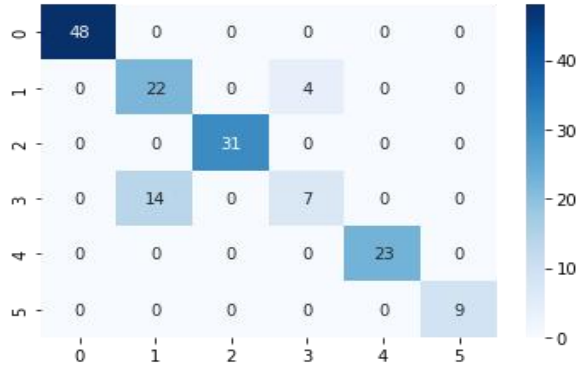


Fig 4.10: BAG SVM

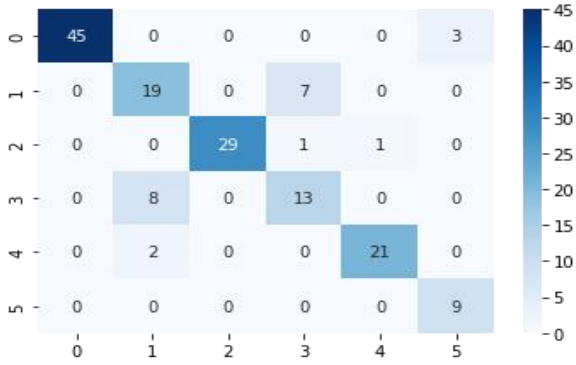


Fig 4.11: Bag DT

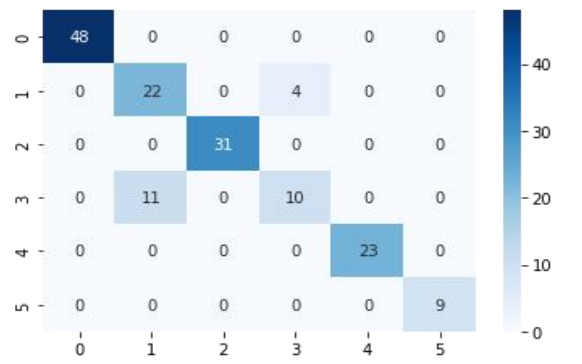


Fig 4.12: Bag MLP

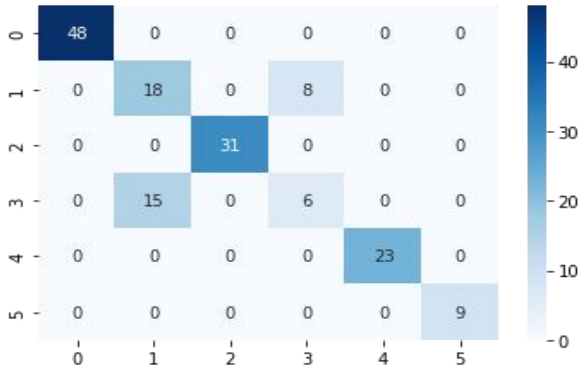


Fig 4.13: Bag KNN

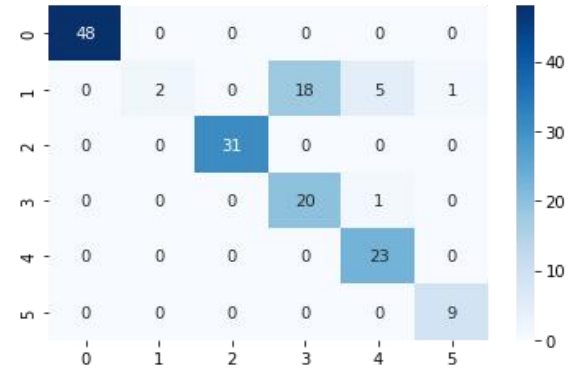


Fig 4.14: Bag GNB

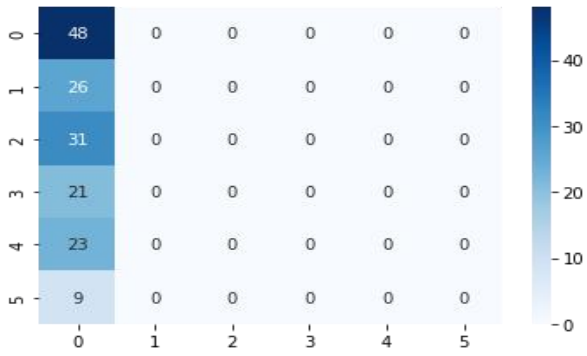


Fig 4.15: Boost SVM

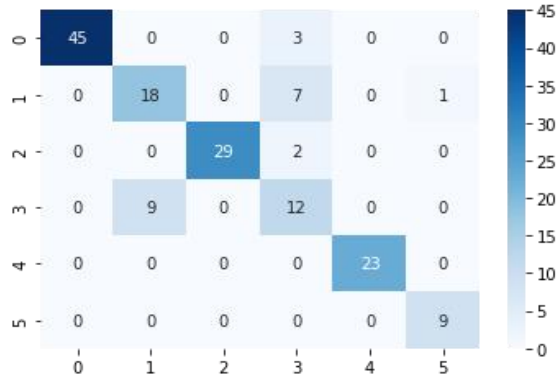


Fig 4.16: Boost DT

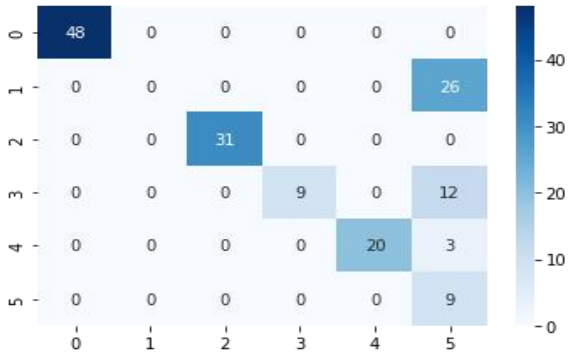


Fig 4.17: Boost GNB

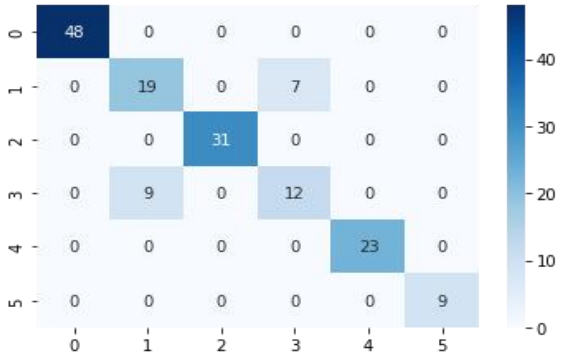


Fig 4.18: Stacking (All Models)

MODEL PERFORMANCE MATRICES:

Figure 4.19 shows a bar chart of the various classification algorithm accuracies. Figure 4.20 also shows the pie chart to illustrate the model accuracies. Table 4.4 tabulate the accuracies and other metrics adopted in this project. From the table 4.4, Bagging MLP shows 90.51% accuracies as the highest followed by SVM and Stacking methods as the second and third highest respectively. From the result achieved, Bagging shows better performance than others.

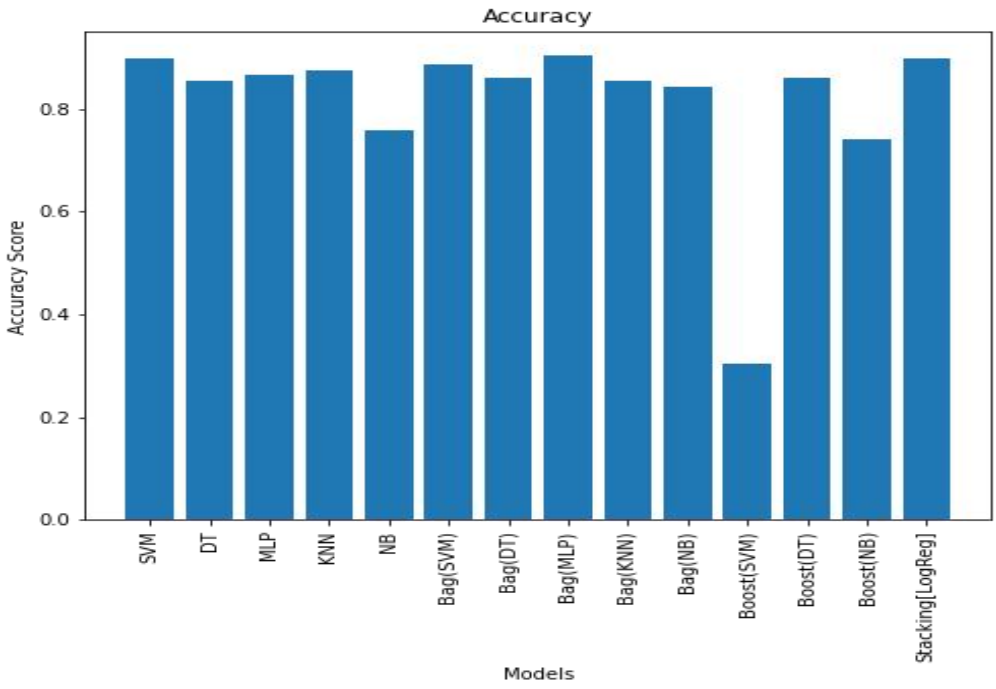


Fig 4.19: A Bar chart of the various classification algorithm accuracies

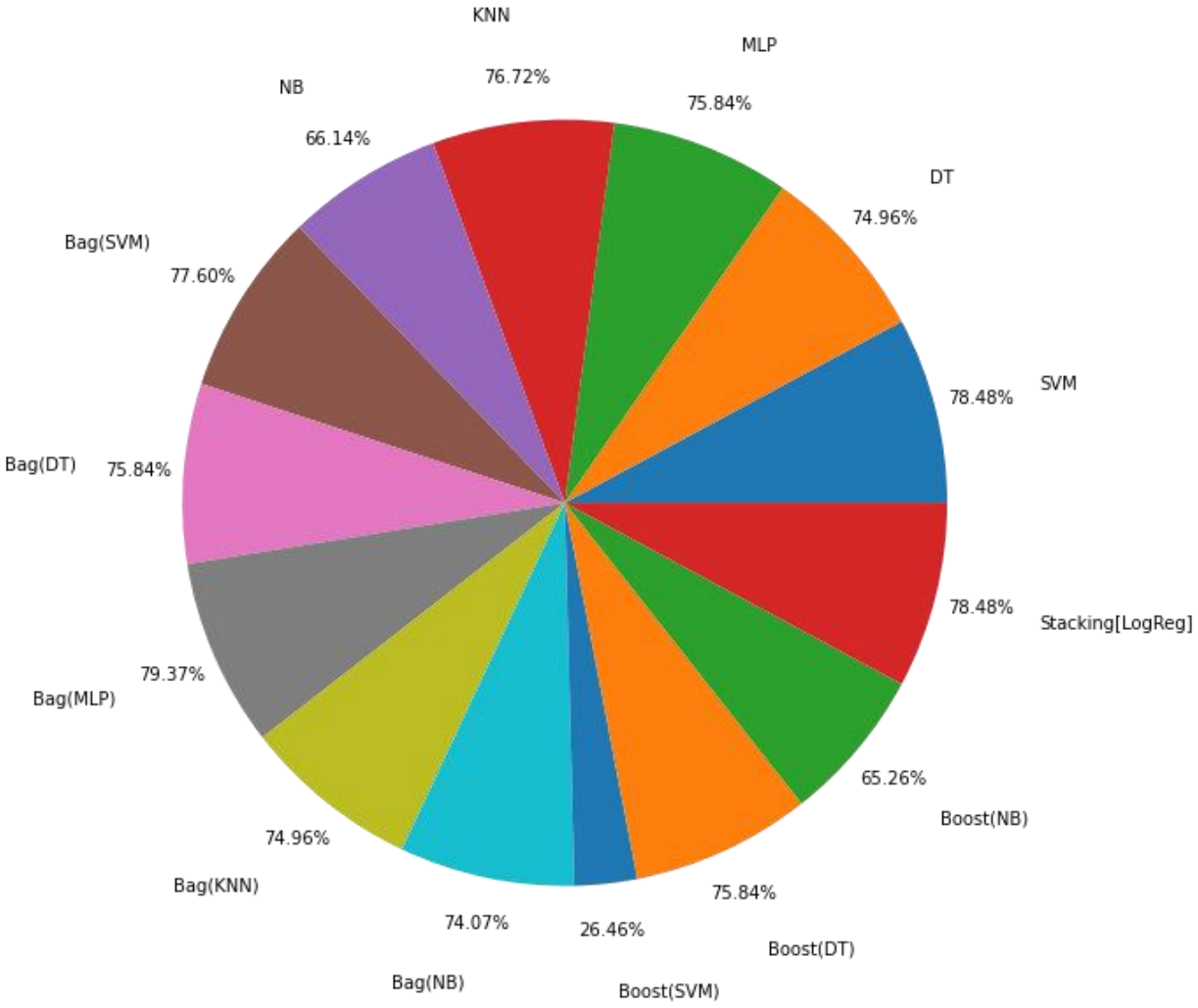


Fig 4.20: The Pie chart to illustrate the model accuracies

Table 4.4: Tableted Accuracies and other metrics adopted in this project

MODELS	ACCURACY	PRECISION	RECALL	SENSITIVITY	SPECIFICITY
SVM	0.898734	0.898734	0.898734	0.8837	0.980121
DT	0.85443	0.85443	0.85443	0.848874	0.972273
MLP	0.867089	0.867089	0.867089	0.833333	0.973485
KNN	0.873418	0.873418	0.873418	0.85348	0.975116
NB	0.759494	0.759494	0.759494	0.738639	0.953659
Bag(SVM)	0.886076	0.886076	0.886076	0.863248	0.977457
Bag(DT)	0.860759	0.860759	0.860759	0.855974	0.973051
Bag(MLP)	0.905063	0.905063	0.905063	0.887057	0.981245
Bag(KNN)	0.85443	0.85443	0.898734	0.82967	0.971328
Bag(NB)	0.841772	0.841772	0.85443	0.838217	0.969576
Boost (SVM)	0.303797	0.303797	0.867089	0.166667	0.833333
Boost(DT)	0.860759	0.860759	0.873418	0.85612	0.972919
Boost(NB)	0.740506	0.740506	0.759494	0.716356	0.954139
Stacking [LogReg]	0.898734	0.898734	0.886076	0.8837	0.980121

CHAPTER FIVE

CONCLUSION AND SUMMARY

5.1 SUMMARY

This study applied an ensemble learning approach to design a predictive model for the classification of erythemato-squamous disease using datasets extracted from the UCI education. Using a hybrid feature selection technique containing four (4) filter based selection methods and an embedded method to select the subset of the most relevant features from the original features. A combination method (voting scheme) to combine the prediction of the different models. Several machine learning classifiers were applied and three (3) ensemble methods were used which are stacking, bagging and boosting.

Using python programming language, the ensemble method combined the multiple classifiers to generate a strong model to improve the accuracy performance for the classification of skin disease datasets when fed with unknown datasets.

5.2 CONCLUSION

This project proves that using ensemble methods to predict erthemato squamous disease can address some of our challenges. The designed model has proven accurate especially when using ensemble methods which gave the highest accuracy.

REFERENCES

Accounting Capital. (2022)What is Gaining Ratio? https://www.accountingcapital.com/ratios/gaining-ratio/

Ain, Q.U., Al-Sahaf, H., Xue, B. and Zhang, M. (2020). A Genetic Programming Approach to Feature Construction for Ensemble Learning Skin Cancer Detection. *GECCO '20: Genetic AND Evolutionary Computation Conference. ACM.*

American Academy of Dermatology Association. (2022). Types of Skin Cancer. American Academy of Dermatology Association. https://www.aad.org/public/diseases/skin-cancer/types/common

American Academy of Dermatology Association. (2022). How can I tell if I have Skin Cancer?. https://www.aad.org/public/diseases/skin-cancer/find/know-how

American Academy of Dermatology Association. (2022). Skin Cancer. https://www.aad.org/media/stats-skin-cancer

ASCO foundation. (2002). Skin Cancer (Non – Melanoma): Introduction Cancer.Net Editorial Board. https://www.cancer.net/cancer-types/skin-cancer-non-melanoma/introduction

Badrinath, N., Gopinath, G., Ravichandran, K.S. and Girish Soundhar, R. (2015). Estimation of automatic detection of erythemato-squamous diseases through AdaBoost and its hybrid classifiers. *Springer: Cross Mark*. pp 471 – 488.

Bose, P., Bandyopadhyay, S., Bhaumik, A. and Poddar, S. (2021). Skin Disease Detection: Machine Learning VS Deep Learning. Preprints.

Brownlee, J. (2019). How to choose a Feature Selection Method for machine learning. *Machine Learning Mastery.*

Chauhan, N.S. (2022). Decision Tree Algorithm, Explained. Machine Learning. KDnuggets. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook, pp. 853–867. Springer US.

Cleveland Clinic. (2021). Skin Cancer. https://my.clevelandclinic.org/health/diseases/15818-skin-cancer

Colin, M. (2022). Multilayer Perception. https://en.wikipedia.org/wiki/Multilayer_perception

Early Detection of Melanoma. (n.d.) https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTacibShbVKOLDR4geK2LhrwVYXhZiQcHKNtg&usqp=CAU

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), pp. 861–874.

Garg, R. (2018). 7 types of classification Algorithms. Developers Corner. AIM. https://analysticsindiamag.com/7-types-of-classification-algorithms/

Gathers, R. C., Feintuch, S., and Wright, S. A. (2022). What is skin cancer?. Healthline. https://www.healthline.com/health/skin-cancer#types

Gathers, R. C., Feintuch, S., and Wright, S. A. (2022). What is skin cancer?. Healthline. https://www.healthline.com/health/skin-cancer

Gathers, R. C., Feintuch, S., and Wright, S. A. (2022). What is skin cancer?. Healthline. https://www.healthline.com/health/skin-cancer#cause

Gupta, A. (2020). Feature Selection Techniques in Machine Learning. *Analytics Vidhya.*

Igodan, C. E., Obe, O., Thompson, A. F-B., and Owolafe, O. (2022) Erythemato Squamous Disease Prediction using Ensemble Multi-feature Selection Approach. International Journal of Computer Science and Information Security (IJCSIS). Vol. 20,No. 2, 95-106.

Jiang, P. (2021). CNN-based Diagnosis System on Skin Cancer using Ensemble Method Weighted by Cubic Precision. *Tech RXIV*.

Kausar, N., Hameed, A., Sattar, M., Ashraf, R., Imran, A.S., Abidin, M.Z.U. and Ali, A. (2021). Multiclass Skin Cancer Classification using Ensemble of Fine-Tuned Deep Learning Models. *Applied Sciences.* **11**: 10593.

Kubat, M. &Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One Sided Selection. Proceedings of the Fourteenth International Conference on Machine Learning, pp 179–186, Nashville, Tennessee. Morgan Kaufmann.

Laurburke. (2022). Logistic Regression. https://en.m.wikipedia.org/wiki/Logistic_Regression

MaterialScientist. (2022). Human Skin. https://en.m.wikipedia.org/wiki/Human_Skin

Mayo Clinic. (2022). Skin Cancer. https://www.mayoclinic.org/diseases-conditions/skin-cancer/symptoms-causes/syc-20377605

Michalska,M. (2021). Overview of Feature Selection Methods used in Malignant Malanoma Diagnostics. *IAPGOS.* PP 32 – 35.

Oliveira, R.B., Pereira, A.S. and Tavares, J.M.R.S. (2017). Skin Lesion Computational Diagnosis of Dermoscopic Images: Ensemble models based on Input Feature Manipulation. *Computer Methods and Programs in Biomedicine.*

OncoBeta. (n.d.). The layers of Human Skin. Epidermal Radioisotope Therapy. https://www.oncobeta.com/your-health/nmsc-info/the-skin

Pace, C. (2017). What is Machine Learning? Definition and Example. *Recorded Future.* https://www.recordedfuture.com/machine-learning-definition.

Pirehelokan. (2022). Naïve Bayes Classifier. https://en.m.wikipedia.org/wiki/Naive_Bayes_Classifier

Rustam, Z. and Kharis, S.A.A. (2020). Comparison of Support Vector Machine Recursive Feature Eliination and Kernel Function as Feature Selection using Support Vector Machine for lung cancer classification. *Journal of Physics: Conference Series*.

Thomas., and Hart, P. (1967) Nearest neighbor pattern classification. Information Theory. 21-27. https://www.cs.cornell.edu/courses/cs4780/2018Fa/lectures/lecturenote02_kNN.html

Tuysuzoglu, G., Birant, D., and Pala, A. (2018). Ensemble Methods in Environmental Data Mining. IntechOpen. http://dx.doi.org/10.5772/intechopen.74393

Verma A.K. AND Pal, S. (2019). Prediction of Skin Disease with Three Different Feature Selection Techniques using Stacking Ensemble Method. *Applied Biochemistry and Biotechnology.*

Verma, A.K., Pal, S. and Kumar, S. (2019). Classification of Skin Disease using Ensemble Data Mining Techniques. *Asian Pacific Journal of Cancer Prevention.* Vol 20. pp 1887-1894.

Verma, A.K., Pal, S. and Kumar, S. (2019). Comparison of Skin Disease prediction by feature selection using ensemble data mining techniques. *Information Medical Unlocked.* IMU 100202.

Verma, A.K., Pal, S. and Kumar, S. (2019). Prediction of Skin Diseases using Ensemble Data Mining Techniques and Feature Selection Method – A comparative study. *Applied Biochemistry and Biotechnology.*

Verma, A.K., Pal, S. and Tiwari, B.B. (2020). Skin Disease prediction using ensemble methods and a new hybrid feature selection technique. *Journal of Computer Science*.

Villacampa, O. (2015). Feature Selection and Classification Methods for Decision Making: A Comparative Analysis. PhD. Thesis.

Wolff, R. (2020, August 26). 5 types of Classification Algorithms in Machine Learning. Monkeylearn. https://monkeylearn.com/blog/classification-algorithms/

World Cancer Research Fund International. (n.d.). Skin Cancer. https://www.wcrf.org/diet-activity-and-cancer/cancer-types/skin-cancer/

Zhou, Z.H. (2012). Ensemble Methods Foundations and Algorithms. Chapman & Hall / CRC.