CLASSIFICATION OF SKIN DISEASES USING MACHINE LEARNING ALGORITHMS

A Project report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

Submitted by

P.P. ANJANA SRIYA (319126512169) **G. TEJA VINAY (319126512144)**
R. NIKHILA (319126512175) **M.SAI CHANDH (319126512163)**

Under the guidance of

Mr. D. Anil Prasad

Assistant Professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES
(UGC AUTONOMOUS)
(*Permanently Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with 'A' Grade*)
Sangivalasa, Bheemili mandal, Visakhapatnam dist.,
(A.P) (2022-2023)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES (UGC AUTONOMOUS)
(Permanently Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with 'A' Grade)
Sangivalasa, Bheemili mandal, Visakhapatnam dist. (A.P)

PRAGNANAM BRAHMA

ANITS

CERTIFICATE

This is to certify that the project report entitled

"CLASSIFICATION OF SKIN DISEASES USING MACHINE LEARNING ALGORITHMS"

submitted by **P.P. Anjana Sriya (319126512169), G. Teja Vinay (319126512144), R. Nikhila (319126512175), M. Sai Chandh (319126512163)** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Electronics & Communication Engineering** of Andhra Pradesh, Vishakapatnam is a record of bonafide work carried out under my guidance and supervision.

Project Guide

Mr. D. Anil Prasad
Assistant Professor
Department of E.C.E
ANITS

Assistant Professor
Department of E.C.E.
Anil Neerukonda
Institute of Technology & Sciences
Sangivalasa, Visakhapatnam-531 162

Head of the Department

Dr. B. Jagadeesh
Professor & HOD
Department of E.C.E
ANITS

Head of the Department
Department of E C E
Anil Neerukonda Institute of Technology & Sciences
Sangivalasa - 531 162

ACKNOWLEDGEMENT

We would like to express our deep gratitude to our project guide **Mr. D. Anil Prasad,** Assistant professor, Department of Electronics and Communication Engineering, ANITS, for his/her guidance with unsurpassed knowledge and immense encouragement. We are grateful to **Prof. B. Jagadeesh**, Head of the Department, Electronics and Communication Engineering, for providing us with the required facilities for the completion of the project work.

We are very much thankful to the **Principal and Management, ANITS, Sangivalasa,** for their encouragement and cooperation to carry out this work. We express our thanks to all **teaching faculty** of Department of ECE, whose suggestions during reviews helped us in accomplishment of our project. We would like to thank **all non- teaching staff** of the Department of ECE, ANITS for providing great assistance in accomplishment of our project.

We would like to thank our parents, friends, and classmates for their encouragement throughout our project period. At last, but not the least, we thank everyone for supporting us directly or indirectly in completing this project successfully.

PROJECT STUDENTS
P.P. ANJANA SRIYA (319126512169)
G. TEJA VINAY (319126512144)
R. NIKHILA (319126512175)
M. SAI CHANDH (319126512163)

CONTENTS

ABSTRACT

Machine learning algorithms were used along with image processing techniques for the detection of skin diseases. Compared to machine learning, the dermatological technique for identifying the type of skin illness is quite expensive. Skin conditions can be surprising and difficult to diagnose. The color of healthy skin differs from that of diseased skin.

Aim of the project is to classify the type of disease using machine learning. In order to decide, these algorithms employ feature values from effected images as input. Three parts make up the process of determining the type of skin disease: feature extraction, training, and testing.

The objective is to increase the accuracy of diagnosis for various types of skin diseases. Here we used herpes, keratosis, eczema, and acne as the four diseases in this study. The process makes use of machine learning technology to train itself with the various skin images. Three important features in image classification are texture, color, shape, and combination of these. In this project work, features such as entropy, variance, contrast, and energy are used to build machine learning algorithm such as logistic regression, Support Vector Machine, and Decision Tree. Accuracy is used to test the performance of the chosen algorithms.

LIST OF SYMOBLS

Y	Dependent variable
x	Independent variable
a_0	y-intercept line
a_1	Slope line
$G(i, j)$	Distribution probability of Gray-level difference between adjacent pixels
L	Number of pixels
x	Pixel value
$\bar{x}$	Mean of pixel values

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

DT	Decision Tree
SVM	Support Vector Machine
LDA	Linear Discriminant Analysis
2-D	Two Dimension
RBF	Radial Basis Function
CART	Classification and Regression Tree
ID3	Iterative Dichotomiser 3
NDA	Normal Discriminant Analysis
DFA	Discriminant Function Analysis
KNN	K-Nearest Neighbour
GLCM	Gray Level Co-occurrence Matrix
RGB	Red, Green, and Blue
HSV	Hue, Saturation, and Value
HSB	Hue, Saturation, and Brightness
MATLAB	Matrix laboratory
PIP	Performance Improvement Plan
IDE	Integrated Development Environment
GUI	Graphical User Interface
AI	Artificial Intelligence

CHAPTER 1
INTRODUCTION

In the present world, skin diseases refer to a wide range of medical conditions that affect the skin, hair, and nails. These diseases can be caused by a variety of factors, including genetics, infections, allergies, autoimmune disorders, and environmental factors such as sun exposure. Some skin diseases are temporary and easily treatable, while others are chronic and require ongoing management. Many skin diseases cause discomfort, pain, and cosmetic concerns, which can affect a person's quality of life. Skin diseases can affect people of all ages and skin types, although some conditions are more common in certain populations. Despite the fact that technology has advanced quickly in the information age, seeing a dermatologist is highly expensive nowadays. These days, face identification and segmentation for skin-damaged areas are only a few applications where image processing and machine learning techniques are applied. In recent days, skin conditions have spread around the globe. Due to poor living conditions, unsanitary surroundings, environmental factors, and individual hereditary conditions, skin disorders have become a significant and alarming societal concern in recent years. Many researchers have used the extraction of features from skin photographs for the identification of skin diseases. In the research publications, they identified and categorized skin diseases based on color and texture [2][3]. Due to the many different skin tones, the existence of hair on the skin, and the texture of the skin, it is challenging to diagnose skin diseases [4]. Both texture and colour features were used by K.V. Swamy and B. Divya to identify disorders in this method, while texture features produced results that were more precise [1]. V.B. Kumar, S.S. Kumar, and Varun Saboo found that decision trees and ANN exhibit accuracy levels that are almost on par with KNN in their search to identify the model that offers the best level of accuracy [2]. P.R. Hegde, M.M. Shenoy, and B.H. Shekar compared the accuracy of four machine-learning techniques: Support Vector Machine (SVM), Artificial Neural Network (ANN), Linear Discriminant Analysis (LDA), Naive Bayes (NB), and SVM [3]. Edge detection, equalisation, and colour modification were utilised as image processing techniques by D.M.T. Rasanga, G.K.A.A. Tharushika, Ishara Weerathunge, and Pradeepa Bandara to identify illnesses [4]. According to a method brought out by N.V. Kumar, P.V. Kumar, and K. Pramodh, the accuracy of the model increased with an increase in the amount of data from the dataset, but at the expense that it takes several hours to obtain that precision [5]. Machine learning is used in each specialized subject, and clinical field analysis is a new tactic that has been added to it. Machine learning algorithms produce better classification results. As they had given better results in skin disease detection and classification. While they keep attracting the interest of researchers from increasingly diverse fields. The development of the model for classifying skin diseases and other clinical uses spread widely in the present day. We choose to find a machine-learning approach for classifying skin diseases.

1.1 PROJECT OBJECTIVE

Aim of the project is to classify the type of disease using machine learning. In order to decide, these algorithms employ feature values from effected images as input. Three parts make up the process of determining the type of skin disease: feature extraction, training, and testing.

The objective is to increase the accuracy of diagnosis for various types of skin diseases. Here we used herpes, keratosis, eczema, and acne as the four diseases in this study. The process makes use of machine learning technology to train itself with the various skin images. Three important features in image classification are texture, color, shape, and combination of these. In this project work, features such as entropy, variance, contrast, and energy are used to build machine learning algorithm such as logistic regression, Support Vector Machine, and Decision Tree. Accuracy is used to test the performance of the chosen algorithms.

1.2 PROJECT OUTLINE

The following four chapters of this project report are used to present it. Chapter 2 describes the Methodology and Machine Learning Algorithms. Chapter 3 presents Feature Extraction which is used to extract features from the skin disease images. Chapter 4 explains the Spyder we used. Chapter 5 presents the Results of the model. Finally, conclusions are drawn in chapter 6.

CHAPTER 2
MACHINE LEARNING ALGORITHMS

2.1 Linear regression:

One of the most widely used and fundamental machine learning techniques is linear regression. A predictive analysis is conducted using this statistical technique. For continuous, actual, numeric variables like salary, age, product cost, etc., predictions are made by linear regression. The dependent and independent variables are shown to have a linear relationship via the linear regression technique. The link is predicted by linear regression as a slope or straight line. It is a prediction algorithm to predict or estimate one variable using another. The Prediction of Ratings using Opinions is an example of linear regression. In this case ratings are the dependent variable, whereas Opinions are the independent variable.

$y = a_1 x + a_0$ is the equation for Linear Regression.

Here y and x are the dependent and the independent variables respectively.

$a_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$ $\qquad$ $a_0 = \frac{\sum y - b(\sum x)}{n}$

a_1= Slope line and a_0= y-intercept of line

Figure 2.1 shows how a sloped straight line is produced by linear regression model for representing the relationship among the variables.

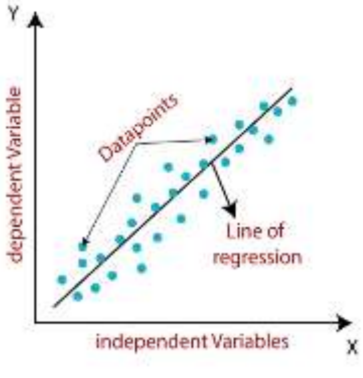


Fig - 2.1: Linear Regression

Types of Linear Regressions:

- **Simple linear regression:** This type of regression consists of only one dependent variable and one independent variable.
- **Multiple linear regression:** In this type of regression one dependent variable or a number of independent variables may both be included.

Advantages of Linear Regression:

- It is efficient to train the data and easy to implement.
- Performing well in linear separable data.

Disadvantages of Linear Regression:

- Noise and overfitting are problems with it.
- Assumption of linearity between variables.

Applications of Linear Regression:

- Prediction of required medicine using no of patients.
- Environmental health – predicting plant growth using an amount of water.

2.2 Logistic regression:

Logistic Regression is a type of supervised learning algorithm. By combining a number of known independent variables, it is utilized to forecast the definite dependent variable. When the dependent variable is categorical, the output is foreseen using logistic regression. With this approach, you may determine if a result is 1 or 0, True or False etc. Similar to linear regression, but with a smaller application to classification techniques, is logistic regression. The algorithm's dependent variable's probability may be calculated. Using this technique, we may produce values ranging from 0 to 1. The logistic regression sigmoid function appears as shown in Fig 2.2.

This figure ranges between 0 and 1. The threshold is set at 0.5. Object belongs to class 0 if the obtained value is less than the threshold value and class 1 otherwise. Logistic regression generally produced two categorical variables like yes or no, 0 or 1 but if the algorithm exceeds two categorical variables the algorithm can have different types named like ordinary, or multinomial.

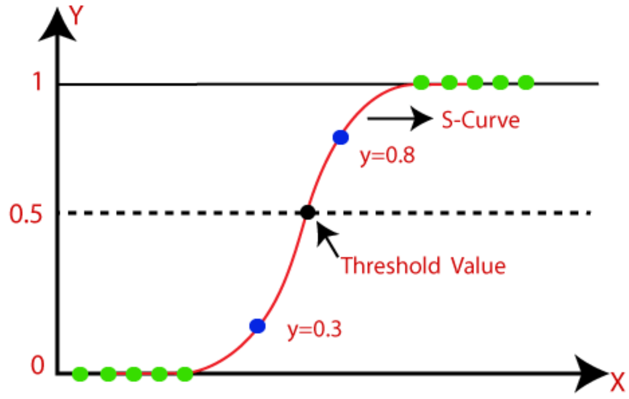


Fig 2.2: Logistic regression

The Equation for Logistic Regression: $y = \frac{1}{1+e^{-x}}$

Where x and y define independent and dependent variables.

We use binary logistic regression basically. Multinomial logistic regression occurs when dependent variable includes three or more choices of unordered categories. Ordinal logistic regression is used when a dependent variable has three or more possible ordered kinds. In logistic regression, an S-shaped curve is used in place of a line. It turns out that the relationship among the variables x and y is non-linear. We can determine the probability of variables using an S-shaped curve.

Advantages of Logistic Regression:

- It doesn't require high computational power to train data·
- It has nice accuracy in simple datasets.
- It can easily be extended to multi classes.

Disadvantages of Logistic Regression:

- We can't handle more categorical variables.
- It is vulnerable to overfitting.

Application of Logistic Regression:

- Probability to get a heart attack or not – Medical research.
- Probability for Transaction is fraud or not - Transactions.
- Probability to check Email spam or not spam – Spam detection.

2.3 Support Vector Machine(SVM):

The Support Vector Machine is also a supervised learning algorithm. SVM's purpose is to use the attributes of the data to build an appropriate hyperplane. It can solve problems of regression and classification too. It gives well conclusion for classification data. Depending on input features SVM can get a model. If features are two the plane of SVM is a line. If features are three the figure shows a 2-D model. SVM classifies the model using hyperplane as we can see it in fig 2.3.

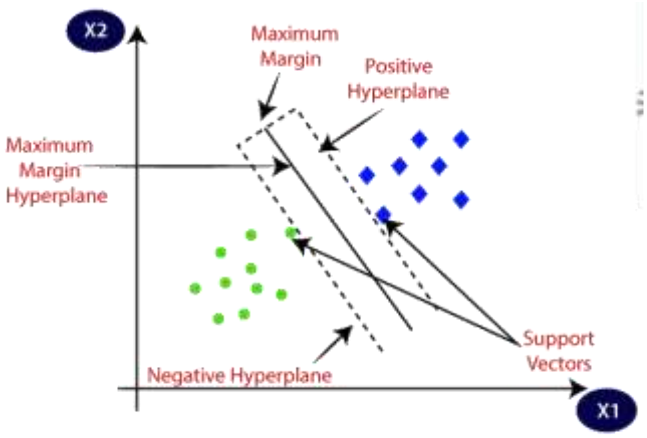


Fig. - 2.3: Support Vector Machine

It has data points called support vectors closest to the plane in the SVM algorithm. Margin can define the SVM either the model is good or bad. The margin is the gap between the data vectors and the plane. A good margin is one that is vast, whereas a bad margin is one that is little.

Types of SVM:

- **Linear SVM:** Linear SVM is used in cases where data can be divided into two classes linearly. This type of data is known as linearly separable data and classifier used here is called as linear SVM classifier.
- **Non-Linear SVM:** The term "non-linear data" refers to data that it is not possible to be classified linearly, while the term "non-linear SVM classifier" refers to the classifier function. Nonlinear SVMs are used to classify data nonlinearly.

In some cases, the algorithm can’t work better, then to improve algorithm efficiency we use kernels. These can make low-dimensionality datasets into high-dimensionality sets.

Linear, polynomial, RBF, sigmoid, etc, are some kernels that give better results than the original one.

Advantages of SVM:

- It can handle many features.
- It has relatively good scaling in high dimensional data.
- Working well with unstructured data.

Disadvantages of SVM:

- It requires a large training time to process.
- It is not suitable for over noise.
- The selection of an appropriate kernel function is challenging.

Application of SVM:

- E-mail classification.
- Face detection – It classifies Faced structure and un face structure then draw a square boundary line to the structure.
- Handwriting recognization – Using SVM classification of character in data.

2.4 Decision Tree (DT):

It is a supervised learning algorithm with a structure like a tree. It has roots and nodes. Nodes of the decision tree have various types like Decision nodes and leaf nodes shown in fig. 2.4

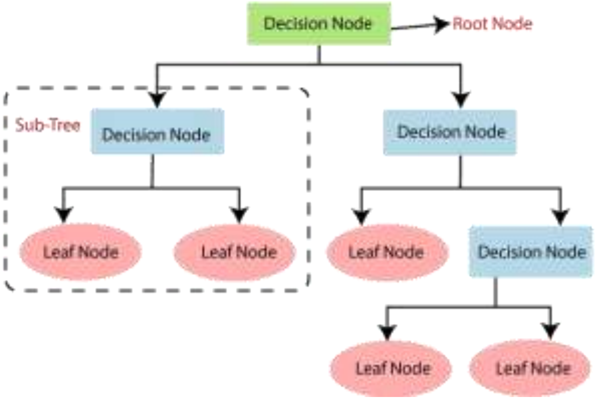


Fig. - 2.4: Decision tree

- Decision trees are often designed to resemble what a person thinks when making a decision, which makes them easier to understand.
- Its logic is simple for understanding due to its tree-like structure.

In a decision tree, the algorithm starts at the Root node and moves upward to identify the dataset's class. This method compares the results of the base attribute and the data (real data) attributes, then follows the branch to move to the next location. Then the value of the next number is compared with the value of the other child nodes, and the process is completed. It continues until it reaches the tree's leaf node, as shown in Fig 2.4.

In the DT, two types of functions used the CART algorithm and ID3 algorithm:

- Gini index comes under the CART algorithm.
- Entropy and Information gain come under the ID3 algorithm.

Advantages of Decision Tree:

- We can add new options for existing trees.
- Easy to understand.
- Less no of data preparation steps to build.

Disadvantages of Decision Tree:

- Its efficiency is less for solving regression problems.
- In the training phase is highly time-consuming.
- We can predict the decision whether a person wants to go for a job or not based on his requirements (Features).

Application of Decision Tree:

- Medical diagnosis.
- Fraud detection.
- Customer segmentation.

2.5 Linear Discriminant Analysis (LDA):

Linear Discriminant Analysis (LDA) is one of the common methods used in machine learning to solve problems that consists of more than two classes. This is an algorithm of supervised machine learning. Its form is a linear discriminant analysis used to separate the dataset into classes by finding linear combinations. Basically, it is a dimensionality reduction approach. Given that certain high-dimensional data may contain noise and redundant data, we limit the number of variables in this dimensionality reduction to convert high-dimensional data into low-dimensional data. In this technique, we separate the dataset into two classes generally later we apply multi-classification as shown in Fig.2.5.

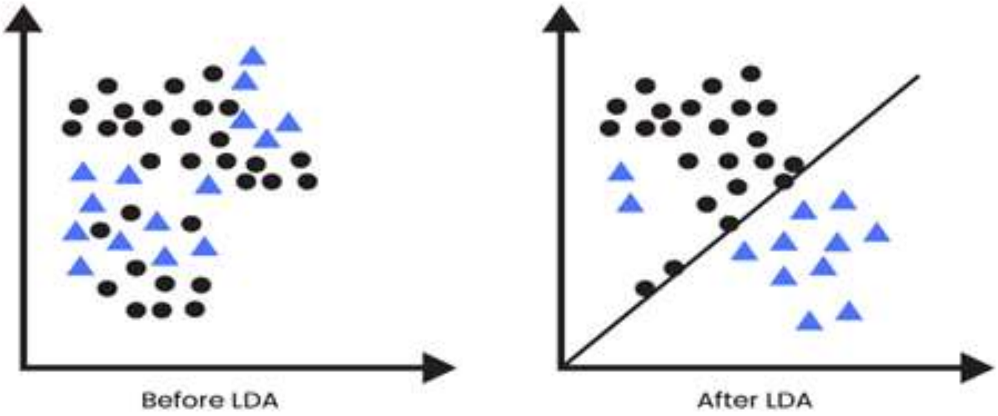


Fig. - 2.5: Linear Discriminant Analysis

Advantages of LDA:

- It is a straightforward and effective approach for computing.
- Even when there are many more characteristics than training examples, it can still perform well.
- It can manage data that has multicollinearity (correlation between characteristics).

Disadvantages of LDA:

- It makes the assumption that the data can be separated linearly, which may not be true for all datasets.
- In feature spaces with several dimensions, it might not function properly.

Application of LDA:

- Medical field - Classification is done based on mild, moderate, and severe diseases using fewer parameters.

- Facial recognition - the removal of some features from a picture before categorization.

2.6 K-Means Clustering Algorithm:

In machine learning or data science, clustering problems uses unsupervised learning method. Clustering is an unsupervised learning method which is used to solve problems in data science or machine learning. K-means clustering is used to solve this kind of problems. Uses an iterative process to split unsigned data into k unique groups, each a group in common with the others. No training needed, this is a simple way to independently identify groups in anonymous data and allows us to aggregate data into different groups. Clustering based on centroid as shown in Fig 2.6.

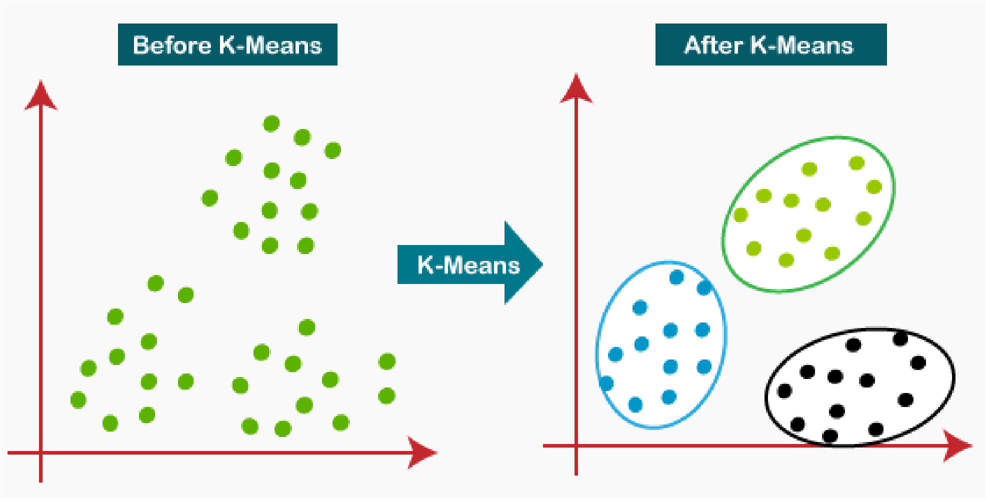


Fig. - 2.6: K-means clustering Algorithm

Since this method is center-based, each cluster has a center point assigned to it. The main purpose of this method is to decrease the distance between each data source and its related groups. The algorithm considers unlabeled data as input, divides it into K groups, and repeats the process until there are no groups left. In this method, the value of K must first be determined. The two main functions of k-means clustering method are:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to

the particular k-center, create a cluster.

Advantages of K-means:

- It can be applied to large datasets and is easy to implement
- When there are many variables, K-means performs quickly.

Disadvantages of K-means:

- It makes the assumption that the data can be separated linearly, which may not be true for all datasets.
- In feature spaces with several dimensions, it might not function properly.

Application of K-means:

- Market Research.
- Market Segmentation.

2.7 K- Nearest Neighbors:

K-Neighbors is one of the simple machine learning methods and is a supervised learning algorithm. The K-NN method, assuming that new and old events are comparable, places new events into categories mostly similar to the existing ones. The K-NN method classifies new content by similarity and records all previous information in the process. This shows that new information can be classified reliably and quickly using the K-NN method. Although the K-NN technique can be used to solve regression and classification problems, it is mostly used for classification types of problems. As K-NN is a non-parametric method, it doesn't make any predictions about the underlying data. This technique is sometimes referred to as lazy learning because the training data is saved rather than immediately learned. Instead, it uses the dataset to do the work when splitting the data. KNN categorizes the new data into clusters close to the training data, keeping the data only during the training period.

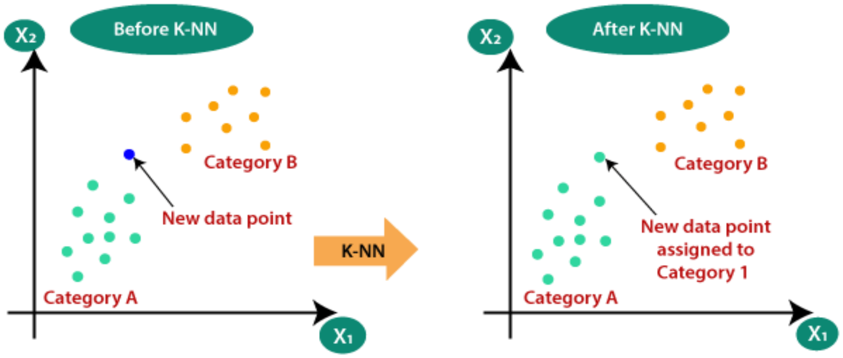


Fig. - 2.7: K-Nearest Neighbors

Advantages of KNN:

- It doesn’t contain highly problematic equations.
- It is used for classification and regression.
- It can give good accuracy.

Disadvantages of KNN:

- It requires large storage capability.
- Increasing noise in training data will lead to low accuracy.
- Irrelevant attributes also can cause low accuracy.

Application of KNN:

- Auto-mobile company- A company wants to know about their prototype and how it is different from other company prototypes so using KNN they know similarities of features in theirs and others.
- Education field – Classification of students based on behavior and attendance so prediction to pass or fail is easy.

CHAPTER 3
FEATURE EXTRACTION

3.1 Introduction:

In this chapter of feature extraction, we will learn about feature selection, problems in feature selection, different type of features, and extracting the features. The classification of skin lesions proceeds to the feature extraction stage, which further extracts the visual appeal of the image and the color characteristics of the skin lesions. In the context of image processing, feature extraction is a technique that reduces a vast number of redundant data into a smaller set of features. Feature extraction is the process of creating a set from the input data. In this project, we have used texture features for classifying skin diseases. The texture features used in this work to classify skin diseases are energy, variance, entropy, and contrast.

So, it will be easier when you want to deal with it. The most important feature of these large files is that they have a large number of variables. These changes require significant computing resources to make them happen. These features are traceable, but still accurate and specific for describing real-world data. When you need to reduce resource usage without losing crucial or pertinent data and have enormous files, extracting features can be quite helpful. The amount of redundant data in a dataset can be reduced with the help of feature extraction. Finally, recovering data speeds up learning and generalization in the machine learning process and enables the construction of models with less computational power.

3.2 Features:

A good feature set contains discrete information that can distinguish one item from other items. It should be as strong as possible against generating different feature values for objects of same class. The selection process should be a small enough that it is useful to distinguish between different models, but similar models in same class. Features can be of four types: Texture features, Color features, Geometric features, and Statistical features.

Here, we have used two features for the feature extraction process and they are:

1. Texture feature
2. Color feature

3.3 Texture Features:

The texture feature is a useful characterization for an image, where pixel properties are used to measure the color in the image while the group of pixels is used to measure a texture.

For many types of ubiquitous images, texture is the most important. Texture is defined as the representation of natural objects by the human eye. It is easy to know everyone, but it is difficult to determine the texture in the matrix, but more and better in the matrix it is revealed in the area where the texture or softness is analyzed. Texture represents the difference at each level, measuring the properties of each surface such as smoothness, roughness, and regularity in different directions. Images are divided into regions of interest and then classified using the texture feature. The spatial distribution of colours or intensities in an image can be learned from the texture. The spatial arrangement of intensity levels within a neighborhood defines the texture. The texture is a repeating pattern of local variations in image intensity.

For example, an image has a 50% black and 50% white distribution of pixels.

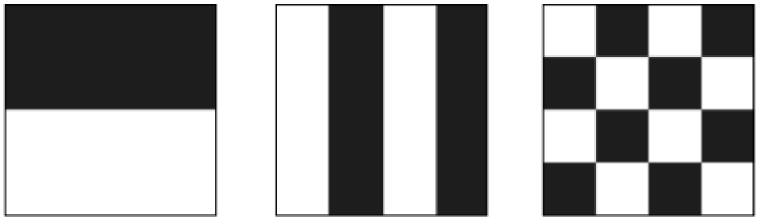


Fig – 3.1: Distribution of pixels

In the above figure-3.1 all the three images have same intensity distribution, but have different textures. There are various texture features such as entropy, homogeneity, range, grey level co-occurrence matrix (GLCM), maximum probability, moments, correlation, and so on.

3.4 Color features:

Since color is an important feature that people perceive when visualizing, it is calculated for feature extraction. Color is the best feature of all visual systems and is widely used in image acquisition systems. Color is usually defined in the 3D color space. They can be Hue, Saturation, and Brightness (HSB), Hue, Saturation, and Value (HSV), or Red, Green, and Blue (RGB) (HSB). Hue, Saturation, and Value (HSV) and Hue, Saturation, and Brightness (HSB), the final two, are based on colour, saturation, and brightness as seen by humans. In our work, we have used four features they are energy, variance, entropy, and contrast.

3.5 Types of features:

3.5.1 Energy:

Energy is used to measure uniformity. The results are same for all combinations resulting in small energy profiles; on the contrary, high energy might be required where results are not balanced. It provides information on image homogeneity. Energy has a low value when the probabilities of the grey level pairs are similar and high values otherwise. The term energy describes the local changes of a certain quality of the image.

Some examples of quality are like brightness and intensity. Energy is specifically important for applications such as image compression. The reason for this lies in the fact that areas with a lot of energy contain a lot of information. Energy could be simply the sum up of all the gray levels.

$$Energy\ = \sum_{i}^{L-1} \sum_{j}^{L-1} G^2(i,j)$$

3.5.2 Entropy:

Entropy is a measure of information content. It determines the intensity distribution's randomness. Entropy is a measurement of the degree of disorder or randomness present in an image. The more random the image is, the higher the entropy. Entropy mainly reflects on the non-uniformity and complexity of image texture.

$$Entropy = -\sum_{i}^{L-1}\sum_{j}^{L-1}\left[G(i,j)\ log\ log\ \left(G(i,j)\right)\right]$$

Such a matrix represents an image where the distance vector has no preferred grey levels. Entropy is lowest when the entries in G I j] are unequal and largest when all entries have the same magnitude.

3.5.3 Variance:

The total of the squared differences between the centre pixel's intensity and its neighbours is known as variance. Variance in image processing refers to a measurement of how much the pixel values in an image deviate from the average value of all the pixels present. The level of details in the image will be higher the if the variance value is high. The variance of a picture can be used to analyse distorted images as well as to quantify the intensity of the image.

$$Variance = \frac{\sum(x-\bar{x})^2}{N}$$

3.5.4 Contrast:

The local variation that can be found in an image is measured using contrast. The difference in colour or grayscale between various images in analog and digital images is referred to as contrast. Images with higher contrast typically exhibit a wider range of colour or grayscale. After an image has been captured with a digital camera or converted from an analog to digital format, its brightness (or brightness) is measured. The difference between the image's greatest and lowest intensity values is known as contrast. This makes it simple to compute from its histogram. Example: If you have a white image, the contrast is 255-255=0 because the lowest and highest values are both 255.

$$Contrast = \sum_{i}^{L-1}\sum_{j}^{L-1}(i-j)^2G(i,j)$$

If there is huge amount of variation among an image the G [i, j]'s will be deviated away from the main diagonal, leading to high contrast value.

3.6 Feature Selection:

Feature extraction and feature selection are the two methods available for obtaining a subset of features. In contrast to feature extraction, which extracts different features, feature selection selects a subset of the initial feature collection. By removing features with poor or unexpected properties, feature selection aims to pick a subset of input variables while maintaining or increasing classification accuracy. John et al. claim that feature relevance can either be strong or weak. If a feature cannot be removed from the feature set without affecting classification accuracy, it is considered to be strongly relevant. When attributes are just slightly significant, it is impossible to reduce the cluster size without also affecting classification accuracy. Variable selection and feature selection are similar terms.

It adopts data before data mining technology is used to reduce data by removing certain values. This technique improves prediction performance, reduces algorithm training time, and provides better data visualization. Special selection medical, assembly process, drawing process, etc. It has many registration applications. In machine learning methods, feature selection methods are basically, divided into three categories: (1) filter method, (2) wrapper method, and (3) embedded method.

For classification issues, especially when employed for handwriting identification, choosing the significant features is a vital stage because:

1. It is necessary to find all possible feature subsets that can be formed from initial values which leads to time consumption,
2. Every feature is meaningful for at least some discriminations, and
3. Variations within the interclass and intraclass are not that much high. Beyond a certain level, the addition of additional features results in a worse performance rather than a better performance.

3.7 Feature extraction:

After selecting the required features for our work, the next stage for the classification is feature extraction in which image texture and color features of the skin images are extracted. Feature extraction is used mainly to increase the accuracy of disease-type detection. Three important features in image classification are texture, shape, color, and also a combination of these features. Here, texture and color features are considered for classifying skin diseases.

Feature extraction refers to the process of transforming raw data into a mathematical process while preserving the information in the original database. It gives better results than applying machine learning directly to raw data. Feature extraction redefines the size of redundant data to the reduced length of pseudo features. In feature extraction, the information obtained from the image is transformed and recorded in a specific set of features to be used for further classification. This stage is used for extracting and identifying features that are derived from objects that were segmented from parts of the image or the whole image. In other words, transforming the data that is obtained from the image into the set of features for pattern recognition is called feature extraction. Selecting the appropriate collection of feature extraction to obtain the necessary data to complete the task and obtain reliable information is one of the key difficulties in pattern recognition. Character recognition, reading bank deposit slips, credit card applications, tax forms, data entry, check-to-sort, and many other uses have all made use of feature extraction. Other uses include extracting necessary image values like entropy and variance. Several features, including shape features (such as area, perimeter, solidity, and others), texture features (such as homogeneity, energy, angular second, entropy, contrast, and others), statistical features (such as mean, skewness, and variance), geometrical features (such as perimeter, area, compactness, and symmetry), colour features, and so on, can be extracted from the objects in the image. In skin disease classification we mainly concentrated on texture features as different diseases will have different textures. In classifying the skin disease, the features extraction stage plays an important role in determining the disease type because different diseases have different feature values.

In machine learning, pattern recognition, and image processing, feature extraction begins with an initial set of data metrics and produces results (features) that are intended to be informative, not repetitive, facilitating further studies and extension, and in some cases, the resulting. advancement of human well-being. Feature extraction plays an important role in image processing. Binarization, thresholding, resizing, normalization, etc., into the image structure before the features are obtained. Various image preprocessing is applied. Feature analysis usually involves low-level processing of images. It checks each pixel to see if a property is available for that pixel.

Feature extraction is used to extract the most important information from the raw data and represent that information in a lower ranked field. When the input data for an algorithm is too large to handle and needs to be reprocessed (there is a lot of data but not too much), the input data is converted to a simple model representation of features (also called a feature vector). Converting input data into a feature set is called feature extraction. If the extraction is chosen carefully, using this simple example it is expected to extract relevant information from the input data of the feature set to perform the desired operation, this is perfectly reasonable.

Feature extraction is done after the preprocessing stage of the feature recognition system. The main function of model validation is to take the model idea and expose it as a product. This process can be divided into two phases: special selection and distribution. Certain options are important to the whole process because products may not be able to identify negative features. The criteria for the selection of features given by Lippman are: Features should contain information needed to distinguish between classes without affecting input inconsistency, and should limit the number to allow calculation of the discrimination function and to limit learning requirements. Feature extraction is an important step in the development of the distributed model and it aims to extract important information that characterizes each class. During this process, main features are extracted from objects/characters to create the feature vector. In this process, relevant features are extracted from objects/ alphabets to form feature vectors.

A classifier then uses feature vectors to describe the input with the target output. By looking at these properties, the classifier is easier on different classes because it allows easy differentiation.

3.8 Importance of feature extraction:

Once the preprocessing and desired level of segmentation (lines, words, characters, or symbols) is completed, some feature extraction techniques are applied to the parts to obtain the features, and then classification and postprocessing techniques are used. Special attention should be paid to the subtraction step because it has a clear impact on the recognition works. Special selection for the extraction process is most important to achieve a high acceptance rate. Feature extraction is defined as "extracting information from raw data suitable for classification purposes while reducing model variability with in the class and facilitating model variability between classes". Therefore, great care must be taken in choosing the appropriate feature extraction technique depending on the input to be applied. Considering all these factors, it is necessary to look at various methods of eliminating the consequences that exist in a particular area, including various consequences.

3.9 Extracted Features:

Firstly, we have selected the diseases which we want to classify they are acne, eczema, keratosis, and herpes. Now we have collected the images of the diseased skin using Kaggle datasets.

In our work, we have selected four features named entropy, variance, contrast, and energy for the better classification of skin diseases. Next, we extracted the selected features of the required diseases. The results of the feature extraction were shown below for required disease separately.

Keratosis:

Table -3.1: Extracted features of keratosis

S.no	Entropy	Variance	Contrast	Energy
1	6.855	3426.488	234.843	22.783
2	6.674	1551.767	129.207	21.876
3	6.942	1396.821	97.726	20.316
4	6.546	1544.117	108.891	14.938
5	7.325	1247.297	102.318	37.179
6	7.437	2008.435	147.77	34.494
7	7.24	2025.067	166.076	35.602
8	6.591	936.089	81.235	13.895
…	…	…	…	…
313	5.79	185.873	14.886	6.264

Eczema:

Table -3.2: Extracted features of Eczema

S.no	Entropy	Variance	Contrast	Energy
1	6.514	740.004	12.926	62.28
2	7.181	1019.697	11.811	73.8
3	5.713	157.438	11.896	12.826
4	7.119	431.661	16.434	38.748
5	6.865	336.584	16.69	29.808
6	7.135	342.671	18.095	29.969
7	6.746	278.737	11.408	28.219
8	6.789	288.45	14.997	30.065
…	…	…	…	…
312	7.207	401.525	13.821	37.587

Herpes:

Table -3.3: Extracted features of Herpes

S.no	Entropy	Variance	Contrast	Energy
1	5.923	393.164	9.61	30.134
2	5.904	994.053	15.189	59.552
3	6.646	3272.831	20.443	196.212
4	6.707	2303.712	17.469	137.802
5	6.169	835.873	14.84	57.64
6	6.479	786.353	19.818	60.088
7	6.049	1026.257	12.189	71.399
8	6.107	1116.573	12.686	79.927
…	…	…	…	…
296	6.422	346.315	8.662	24.621

Acne:

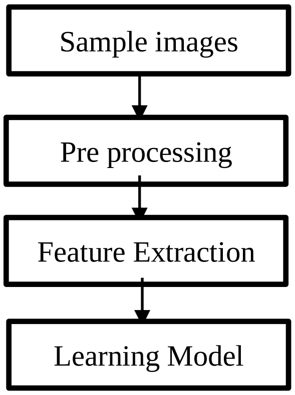
Table -3.4: Extracted features of Acne

S.no	Entropy	Variance	Contrast	Energy
1	6.63	760.666	57.701	12.376
2	6.836	740.343	58.276	12.567
3	6.662	574.406	35.422	9.877
4	7.37	310.911	24.805	11.876
5	7.308	1150.149	77.42	18.178
6	7.07	195.97	16.179	6.676
7	7.027	841.958	51.349	14.066
8	7.358	97.628	8.948	6.292
…	…	…	…	…
289	6.011	1110.64	70.461	20.475

Above tables 3.1, 3.2, 3.3, 3.4 shows the extracted features such as entropy, variance, contrast and energy of diseases named Keratosis, Eczema, Herpes, and Acne respectively.

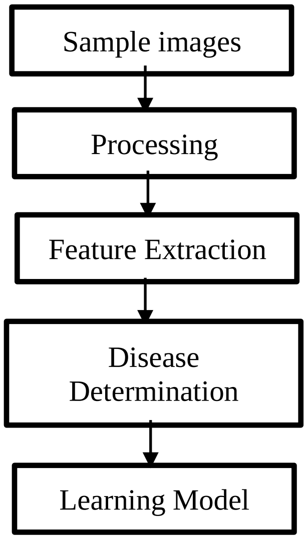
3.10 Methodology

The methodology of Classification of skin diseases consists of two phases namely the Training phase and the Testing phase. Training Phase is a phase where several images of disease classes are obtained, from which image features obtained are stored. The extracted features of different disease classes are used for creating the required learning model using different machine learning algorithms for classifying the skin diseases. In the Testing Phase, few images of disease classes are obtained, from which image features are obtained and are compared with respect to a determined model created using the Training phase. Based on the model given image input is classified accordingly.



.

Training Phase



Testing Phase

CHAPTER 4
THE SPYDER

4.1 Introduction:

For Python scientific programming, Spyder is a well-liked open-source integrated development environment (IDE). It provides a powerful environment for scientific computing and data analysis with a user-friendly interface. Spyder integrates with many of the popular scientific libraries for Python such as NumPy, SciPy, Matplotlib, Pandas, and more, and offers features such as code highlighting, code completion, debugging tools, and an IPython console. Spyder's interface is similar to other popular development environments such as MATLAB and RStudio, making it easy for scientists and researchers to switch to Python for their data analysis needs. It is designed to be flexible and customizable, allowing users to choose their preferred coding style and preferences. Spyder is available on Windows, macOS, and Linux and can be downloaded from the official website or installed using popular package managers such as Anaconda and pip. With its comprehensive features and ease of use, Spyder is an excellent choice for scientific programmers and data analysts who prefer Python.

4.2 Features of Spyder:

Spyder is a powerful Python development environment designed specifically for scientific computing and data analysis. Some of its key features include:

- Interactive console: Spyder includes an IPython console that allows users to run and interact with code in real-time.
- Code completion and introspection: Python code writing and debugging are made simple using Spyder's intelligent code completion and introspection capabilities.
- Code analysis: Spyder includes built-in code analysis tools that help users identify errors, warnings, and code smells in their Python code.
- Variable explorer: Spyder's variable explorer provides an easy way to view, edit, and manipulate variables and data structures in real-time.

- Debugging: Spyder includes a powerful debugger that allows users to step through code, set breakpoints, and view variable values at runtime.
- Profiling: Spyder provides built-in profiling tools that help users identify performance bottlenecks in their code.
- Integrated development environment: Spyder provides a complete development environment with a code editor, file explorer, and project management tools.
- Multi-language support: Spyder supports multiple languages including Python, R, and Octave.
- Customizability: Spyder is highly customizable, allowing users to configure keyboard shortcuts, color schemes, and other preferences to suit their workflow.

Overall, Spyder is a comprehensive Python development environment that provides everything scientists and data analysts need to write, debug, and analyze Python code for scientific computing and data analysis.

4.3 Uses of Spyder:

Spyder is a popular Python development environment that is specifically designed for scientific computing and data analysis. It is widely used in various fields such as:

1. Data analysis: Spyder is extensively used in data analysis tasks as it provides an interactive console, code completion and introspection, a variable explorer, and a powerful debugger. It is a popular option for data analysts and scientists since it interfaces with well-known data analysis libraries like NumPy, Pandas, and Matplotlib.
2. Machine learning: As it supports well-known machine learning frameworks like Scikit-learn and TensorFlow, Spyder is frequently used for machine learning projects. Its interactive console and code completion features make it easy to experiment with different algorithms and models.
3. Scientific computing: Spyder is an excellent choice for scientific computing tasks as it supports scientific libraries such as SciPy and SymPy. It provides a complete

development environment with debugging and profiling tools, making it easy to develop and debug complex scientific code.

4. Education and research: Spyder is a popular choice for educational and research purposes as it is free, open-source, and provides a user-friendly interface. It is frequently employed to teach Python programming, data analysis, and scientific computing in academic contexts.
5. Web development: Spyder can also be used for web development tasks as it supports web frameworks such as Flask and Django. Its code completion and debugging features make it easy to develop and debug web applications.

Overall, Spyder is a versatile Python development environment that is frequently used in a variety of fields, including data analysis, machine learning, scientific computing, education, research, and web development.

4.4 Components of Spyder:

4.4.1 Editor:

The main component of the IDE is Spyder's multilingual Editor pane, where source files can be created, opened, and modified. A wide range of essential functions are available in the Editor, including autocompletion, real-time analysis, syntax highlighting, horizontal and vertical splitting, and many more. Also, it incorporates a number of strong tools for a user-friendly, effective editing experience. The editor in spyder is depicted in Figure 4.1.

Fig - 4.1: Editor in Spyder

4.4.2 IPython Console:

As seen in figure 4.2, the IPython Console enables you to issue commands and work with data inside IPython interpreters. Real-time function calltips, automatic code completion, and full GUI integration with the improved Spyder Debugger.

Fig - 4.2: Console in Spyder

4.4.3 Variable Explorer:

You may view and manage the objects produced by interactively running your code using the variable explorer. It displays the namespace contents of the currently chosen Ipython terminal session, including global objects, variables, class instances, and more, and lets you add, remove, and change their values using a number of GUI-based editors. As seen in figure 4.3, the Variable Explorer provides details on each object's name, size, type, and value. Simply double-click a scalar variable, such as a number, string, or boolean, in the pane and enter its new value.

Fig - 4.3: Variable Explorer in Spyder

4.4.4 Plots:

The static graphics and images produced during your session are displayed in the Plots pane. You can interact with the plots in a variety of ways by seeing those created by the Variable Explorer, the Editor, or the IPython Console. The plots displayed in the Plots pane correspond to the console tab that is presently selected; if you switch consoles, the list of plots presented will adjust as illustrated in figure 4.4.

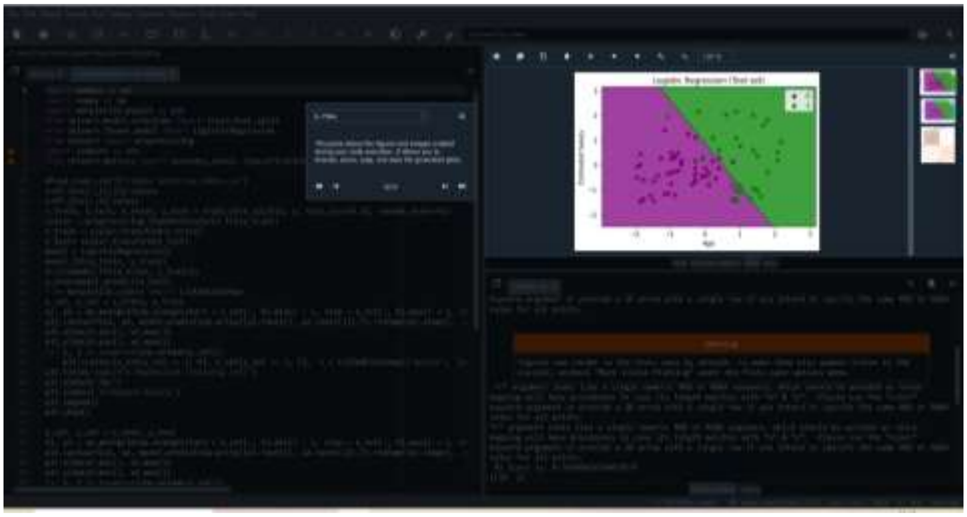


Fig. - 4.4: Plots in Spyder

4.4.5 Files:

The Files pane is a filesystem and directory browser built right into Spyder. You can view and filter files according to their type and extension, open them with the Editor or an external tool, and perform many common operations as shown in figure - 4.5.

Fig. - 4.5: Files in Spyder

4.4.6 Help:

Any object with a docstring, including as modules, classes, functions, and methods, can have extensive documentation that can be found, rendered, and shown using the Help window. As seen in figure 4.6, this enables you to quickly access documentation directly within Spyder without having to pause your process.

Fig. 4.6: Help in Spyder

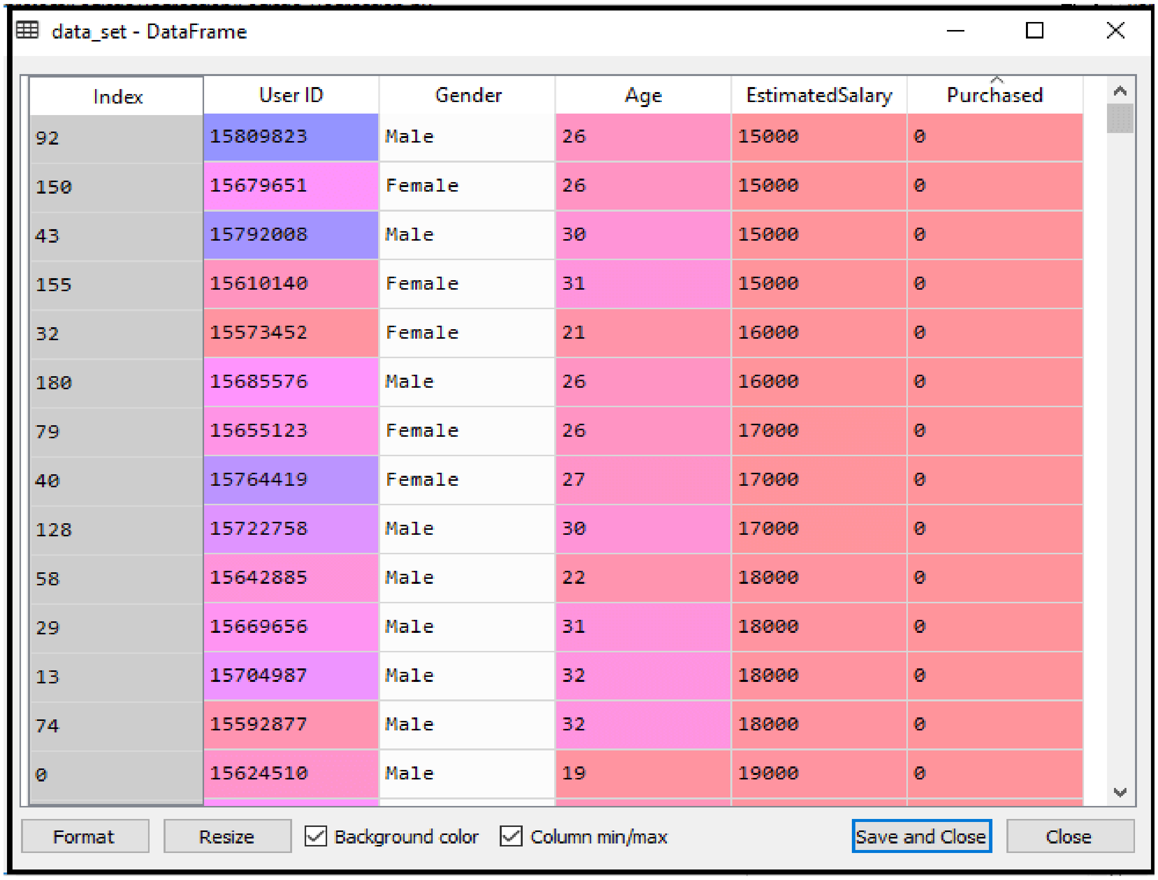
4.5 Machine Learning using Spyder (Python):

A subfield of artificial intelligence (AI) and computer science called "machine learning" aims to simulate how people learn by using data and algorithms, gradually increasing the accuracy of the results. It enables the user to provide an enormous quantity of data to a computer algorithm, which the computer will then evaluate and utilise to provide suggestions and conclusions based only on the supplied data. Let's use an example to explain it.

Example:

A dataset is provided that includes data on various people collected from social networking sites. A new SUV vehicle was just introduced by an automobile manufacturer. The business thus needed to determine how many consumers in the dataset were interested in buying a car. Using the logistic regression approach, we will create a machine learning model for this issue. The below image displays the dataset. The dataset for this issue has 5 columns, as seen in figure 4.7. It consists of UserID, Gender,

Age, Estimated Salary, and Purchased. With age and salary (Independent variables), we will predict the purchased (Dependent Variable) .



Index	User ID	Gender	Age	EstimatedSalary	Purchased
92	15809823	Male	26	15000	0
150	15679651	Female	26	15000	0
43	15792008	Male	30	15000	0
155	15610140	Female	31	15000	0
32	15573452	Female	21	16000	0
180	15685576	Male	26	16000	0
79	15655123	Female	26	17000	0
40	15764419	Female	27	17000	0
128	15722758	Male	30	17000	0
58	15642885	Male	22	18000	0
29	15669656	Male	31	18000	0
13	15704987	Male	32	18000	0
74	15592877	Male	32	18000	0
0	15624510	Male	19	19000	0

Format Resize ☑ Background color ☑ Column min/max Save and Close Close

Fig. - 4.7: Considered Dataset used for the Example

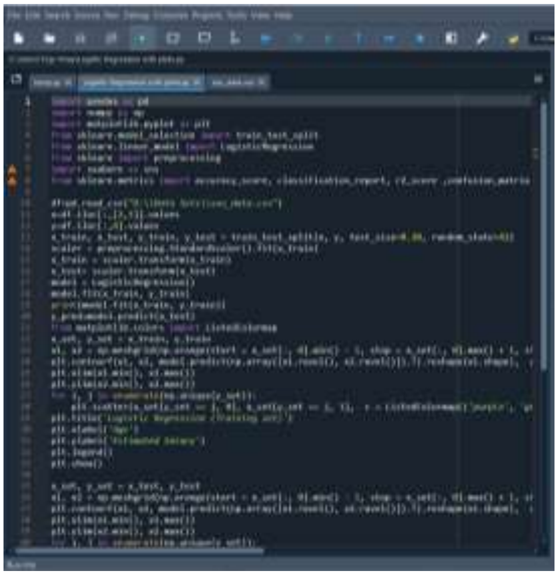
The code will be written in the editor and the code will be executed(run) in the console which is shown in above figures (fig. - 4.8 & 4.9).

In the code model classifies the data using the Logistic Regression

model = LogisticRegression()

model.fit(x_train, y_train)

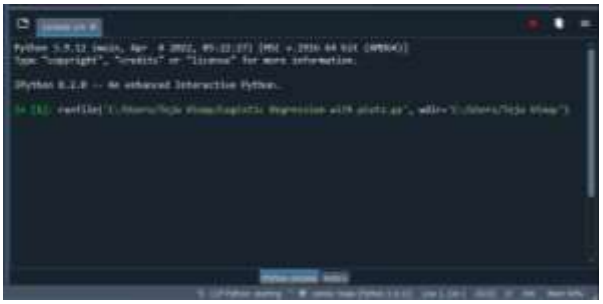
All the variables which are used in the example will be visible in the variable explorer which is shown in figure - 4.10.



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import preprocessing
import seaborn as sns
from sklearn.metrics import accuracy_score, classification_report, r2_score ,confusion_matrix

df=pd.read_csv("[illegible]")
x=df.iloc[:,[2,3]].values
y=df.iloc[:,4].values
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20, random_state=0)
scaler = preprocessing.StandardScaler().fit(x_train)
x_train = scaler.transform(x_train)
x_test= scaler.transform(x_test)
model = LogisticRegression()
model.fit(x_train, y_train)
print(model.fit(x_train, y_train))
y_pred=model.predict(x_test)
from matplotlib.colors import ListedColormap
x_set, y_set = x_train, y_train
x1, x2 = np.meshgrid(np.arange(start = x_set[:, 0].min() - 1, stop = x_set[:, 0].max() + 1, s
plt.contourf(x1, x2, model.predict(np.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),  (
plt.xlim(x1.min(), x1.max())
plt.ylim(x2.min(), x2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(x_set[y_set == j, 0], x_set[y_set == j, 1], c = ListedColormap(('purple', 'gr
plt.title('Logistic Regression (Training set)')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()

x_set, y_set = x_test, y_test
x1, x2 = np.meshgrid(np.arange(start = x_set[:, 0].min() - 1, stop = x_set[:, 0].max() + 1, s
plt.contourf(x1, x2, model.predict(np.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),  (
plt.xlim(x1.min(), x1.max())
plt.ylim(x2.min(), x2.max())
for i, j in enumerate(np.unique(y_set)):
```



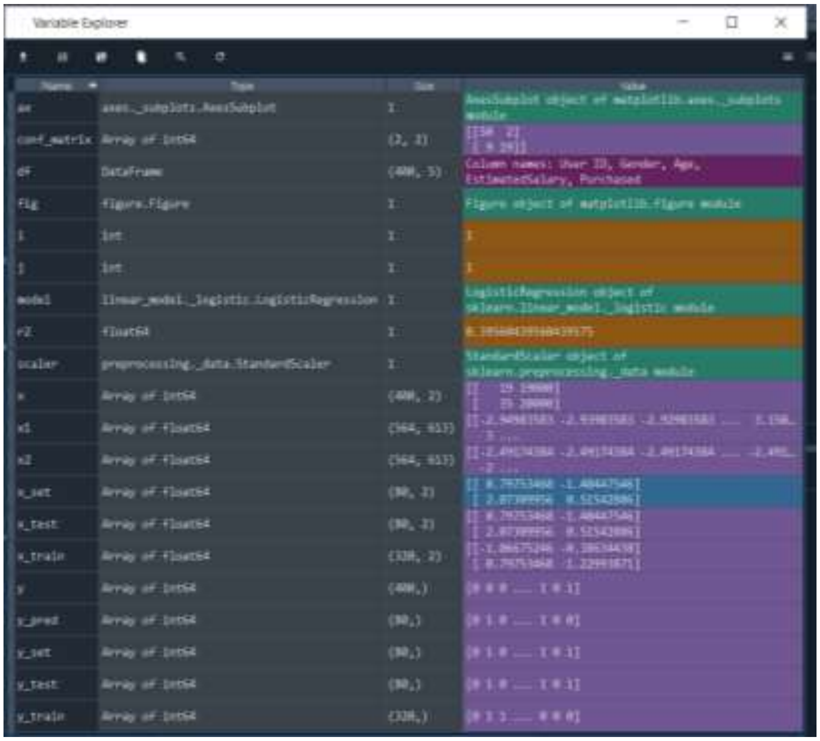
```
Python 3.9.12 (main, Apr  4 2022, 05:22:27) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 8.2.0 -- An enhanced Interactive Python.

In [1]: runfile('[illegible]/Logistic Regression with plots.py', wdir='[illegible]')
```

Fig. - 4.8: Code in the Editor

Fig. - 4.9: Console for executing the code



Variable Explorer

Name	Type	Size	Value
ax	axes._subplots.AxesSubplot	1	AxesSubplot object of matplotlib.axes._subplots module
conf_matrix	Array of int64	(2, 2)	[[50 2] [9 19]]
df	DataFrame	(400, 5)	Column names: User ID, Gender, Age, EstimatedSalary, Purchased
fig	figure.Figure	1	Figure object of matplotlib.figure module
i	int	1	1
j	int	1	1
model	linear_model._logistic.LogisticRegression	1	LogisticRegression object of sklearn.linear_model._logistic module
r2	float64	1	0.[illegible]
scaler	preprocessing._data.StandardScaler	1	StandardScaler object of sklearn.preprocessing._data module
x	Array of int64	(400, 2)	[[19 19000] [35 20000]]
x1	Array of float64	(564, 613)	[[-2.[illegible] -2.[illegible] -2.[illegible] ... 3.[illegible] 3 ...
x2	Array of float64	(564, 613)	[[-2.[illegible] -2.[illegible] -2.[illegible] ... -2.[illegible] -2 ...
x_set	Array of float64	(80, 2)	[[0.79753468 -1.40447546] [2.07399956 0.51542886]]
x_test	Array of float64	(80, 2)	[[0.79753468 -1.40447546] [2.07399956 0.51542886]]
x_train	Array of float64	(320, 2)	[[-1.06675246 -0.38634438] [0.79753468 -1.22993871]]
y	Array of int64	(400,)	[0 0 0 ... 1 0 1]
y_pred	Array of int64	(80,)	[0 1 0 ... 1 0 0]
y_set	Array of int64	(80,)	[0 1 0 ... 1 0 1]
y_test	Array of int64	(80,)	[0 1 0 ... 1 0 1]
y_train	Array of int64	(320,)	[0 1 1 ... 0 0 0]

Fig. - 4.10: Variable Explorer for the example

The Results for the above example can be seen in the Plots mentioned in the below figures (fig. - 4.11 & 4.12).

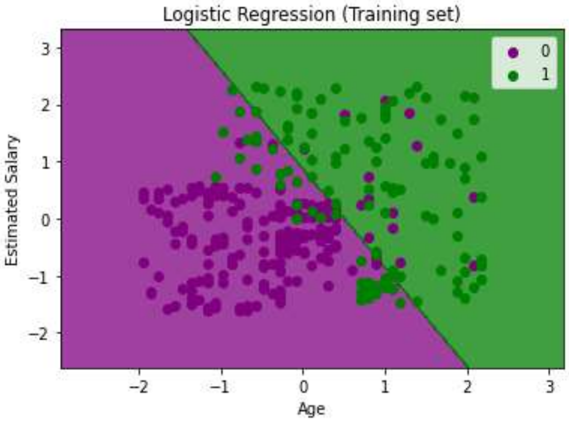


Fig. - 4.11: Training Set

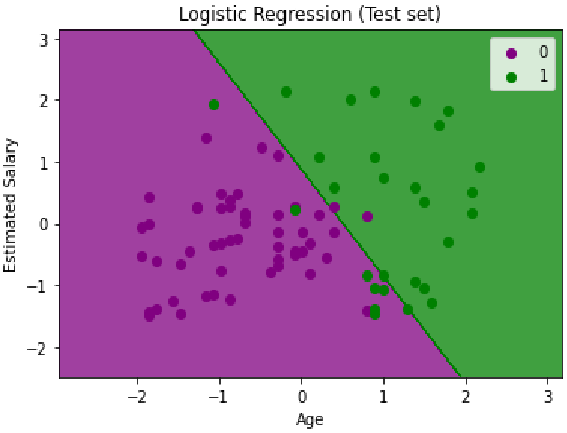
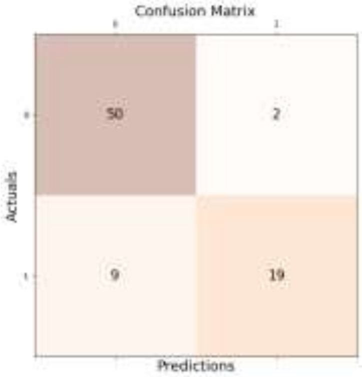


Fig. - 4.12: Testing Set

With the use of the confusion matrix and r2 score, the model's accuracy will be calculated. The confusion matrix is represented graphically in figure 4.13, and the console will display the confusion matrix and r2 score values as seen in figure 4.14 below.



Actuals \ Predictions	0	1
0	50	2
1	9	19

Fig. - 4.13: Confusion Matrix for the example

```
R2 Score is: [illegible]
[[50  2]
 [ 9 19]]
```

Fig. - 4.14: R2 score and Confusion Matrix

Hence, we explained the process of how the code is written and works in the Spyder by showing where the code is written, where the execution is done, representing them in plots, and finally where the output is shown.

CHAPTER 5
RESULTS

5.1 Introduction:

To build a diagnostic system for dermatological diseases firstly, the preparation of the data set, which is the basic building block can be obtained from electronic sites over the Internet. Secondly, the selection of the classification algorithm from the family of supervised machine learning algorithms. Then begins the operation of processing the data and extracting the necessary features that perform a model. In any classification process using machine learning algorithms, the choice of feature to be entered is a very important and key step.

5.2 Skin disease images Dataset:

The proposed algorithms were trained and tested from the features of images which were obtained from Kaggle website. This anonymous dataset was collected from patients suffering from different skin diseases and categorized the images accordingly. These images are obtained from the public access "Dermnet", one of the biggest online dermatology resource designed for providing online medical education. The image is in JPEG format and has three channels RGB. Resolution from image to image and category to category varies but most of the images do not have very high-quality images.

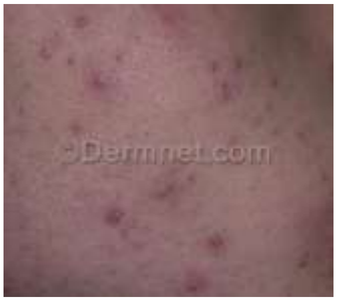


Fig 5.1. Acne

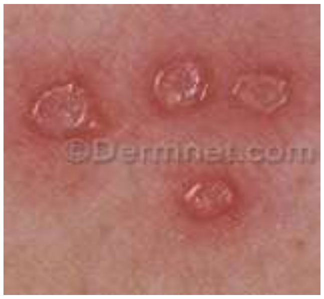


Fig 5.2. Herpes

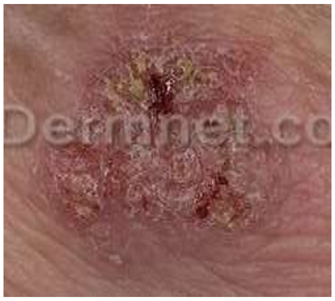


Fig 5.3. Eczema

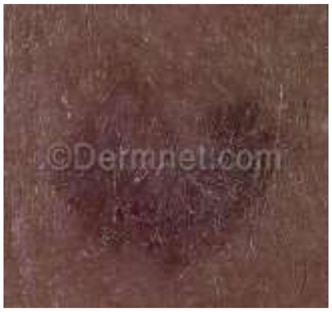


Fig 5.4. Keratosis

The obtained dataset has many disease image collections like Acne, Exanthems, Lupus, Melanoma and such 23 disease sets with total of 19,500 images around, from which we have selected the diseases that occur more often like Acne and Herpes. After downloading the dataset, we have filtered the images which have images with disease shown clearly so that it gives better feature values during extraction. Table 5.1 indicates the count of images considered for different disease classes.

Table 5.1: Count of images considered for each disease set.

Disease	**Count**
Acne	290
Eczema	312
Herpes	297
Keratosis	314
Normal Skin	87

5.3 Image Processing:

Image processing is a technique used to perform few particular operations on a particular image to obtain an enhanced image or for extracting some useful data from it. Image classification was done was based on the extracted features from the images of the dataset of required diseases. For this we used image processing, based methods for purpose of extracting required features from the images.

Image processing methods for image classification can be of Texture features and color features. Features we considered are Entropy, Variance, Energy and Contrast; in which Contrast is a color feature and Entropy, Variance, Energy are Texture Features.

5.3.1 Texture Features:

Entropy:

Entropy, a measure of the quantity of information in an image, is characterized by the equivalent intensities to which individual pixels may adapt.

$$Entropy = -\sum_{i}^{L-1}\sum_{j}^{L-1}\left[G(i,j)\ log\ log\ \left(G(i,j)\right)\right]$$

Variance:

Variance, it is a measure of how the pixel values in an image differ from the average of all the pixel values, gives an idea of how evenly are the pixel values distributed.

$$Variance = \frac{\sum\left(x_i - \underline{x}\right)^2}{N}$$

Energy:

Energy is determined by quadratic sum of the components of the Grey-level co-occurrence matrix in the horizontal and vertical axes gives the energy value, which essentially indicates the texture's thickness.

$$Energy\ = \sum_{i}^{L-1}\sum_{j}^{L-1} G^2(i,j)$$

5.3.2 Colour Features:

Contrast:

Contrast in analogue and digital photos indicates the degree of color or grey scale distinction that is present between different visual aspects.

$$Contrast = \sum_{i}^{L-1} \sum_{j}^{L-1} (i-j)^2 G(i,j)$$

Where '∑' is called sigma (summation)

5.4 Algorithms:

5.4.1 Logistic Regression:

Logistic regression is a Machine Learning algorithm, based on Supervised Learning technique. It is used to predict a definite dependent variable from the given set of independent variables. It uses binary results to indicate how a categorical dependent variable will change over time. Hence, True or False, 0 or 1, Yes or No, etc. are all valid results of a discrete/categorical operation. In spite of this, it produces probabilistic results rather than exact numbers between 0 and 1. Although there are numerous common things between linear regression and logistic regression, the two methodologies use different approaches. Unlike linear regression, which is used to solve regression problems, logistic regression is used to address classification-related concerns. In Logistic Regression, we fit the "S-shaped" logistic function, which forecasts two maximum values, rather than a regression line (0 or 1). Logistic regression produces a curve resembling the "S-curve" because the value must lie within the range of 0 and 1, and it cannot go beyond this limit. A logistic function or sigmoid function are the other names for the S-form curve.

$$log\, log\, \left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Where, $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$

Some applications of Logistic regression are spam mail detection, predicting if a credit card transaction is fraudulent or not fraudulent.

5.4.2 Support Vector Machine (SVM):

Support Vector Machine (SVM) is one of the Machine Learning algorithms, which is a Supervised Learning technique and is used for both classification and regression. Regression problems are best suited for classification. SVM algorithm generates the appropriate decision boundary or line that can categorize the updated data point efficiently in the future by dividing the N-dimensional space into classes, by forming a hyperplane in an N-dimensional space that used for classifying the data vectors distinctly. The dimension of the hyperplane is related to number of features used.

Binary classification is the technique for classifying data when there are only two groups or classes. A Multi-class SVM is used in places where there are groups or classes more than two in the data. Some of the applications of SVM are face detection, intrusion detection, handwriting recognition and gene classification.

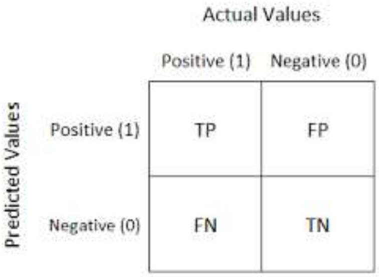
5.4.3 Decision Tree (DT):

Decision Tree, it is a method for supervised learning with a structure like a tree. It has roots and nodes. Nodes of the decision tree have various types like Decision nodes and leaf nodes. Decision trees are often designed to resemble what a person thinks when making a decision, which makes them easier to understand. The logic of the decision tree is easy to understand due to its tree-like structure. In a decision tree, the algorithm starts at the root node and moves upward to identify the dataset's class. This method compares the results of the base attribute and the data (real data) attributes, then follows the branch to move to the next location. Then the value of the next number is compared with the value of the other child nodes.

Expected value (EV) = (First possible outcome x Likelihood of outcome) + (Second possible outcome x Likelihood of outcome) - Cost of each outcome.

5.5 Metrics:

Accuracy and the confusion matrix are used as measures to gauge how well the model's function. Confusion matrix is used for determining the execution of the classification models from the given test dataset. It can only be used if true value for test data is known.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig 5.5: confusion matrix 2×2 representation

As shown in fig 5.5, each column in the matrix refers to events that occurred in the expected class, while each row in the matrix refers to events that occurred in the actual class. The number of actual predicted classes is contained in the diagonal matrix elements. In order to examine how algorithms are being executed, accuracy is used.

Accuracy is a metric that is used to describe how the model works across all the classes. It is useful when all the classes are given same importance. It is a measure of the ratio between the number of correct predictions to the total predictions.

It is used to evaluate how reliable a classification method is.

$$\text{Accuracy}= \frac{(Number\ of\ correct\ predictions\ Accuracy)}{(Total\ Predictions)}$$

$$\text{Accuracy}= \frac{TP+TN}{TP+TN+FP+FN}$$

Where, TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

5.6 Results:

By using different machine learning algorithms highest efficiency obtained was 72.3% by Support Vector Machine (SVM) algorithm. Confusion matrices of these respective algorithms are below.

Table 5.2: Confusion matrix for Logistic Regression algorithm

	Class:1	Class:2	Class:3	Class:4	Class:5	Classification overall
Class:1	33	0	0	10	1	44
Class:2	0	35	12	0	1	48
Class:3	0	18	19	0	1	38
Class:4	10	0	0	37	1	48
Class:5	0	1	0	0	16	17
Truth overall	43	54	31	47	20	195

In table 5.2 indicates the Logistic Regression using sigmoid function where Comparable to a two-layer supervised learning neural network design, this function acts as an activation function for artificial neurons, diagonal elements of obtained matric indicates correct predictions which sum up to give 140 i.e., total 140 predictions were correct out of 195 and giving the total accuracy as 71.79%.

Table 5.3: Confusion matrix for DT algorithm

	Class:1	Class:2	Class:3	Class:4	Class:5	Classification overall
Class:1	28	0	0	16	2	46
Class:2	0	21	19	0	0	40
Class:3	1	12	32	0	1	46
Class:4	20	0	0	34	0	54
Class:5	2	0	0	0	7	9
Truth overall	51	33	51	50	10	195

Table 5.3 indicates the Decision Tree using Gini Criterion where a function that evaluates the effectiveness of the decision tree split, diagonal elements of obtained matric indicates correct predictions which sum up to give 122 i.e., total 122 predictions were correct out of 195 and giving the total accuracy as 62.56%.

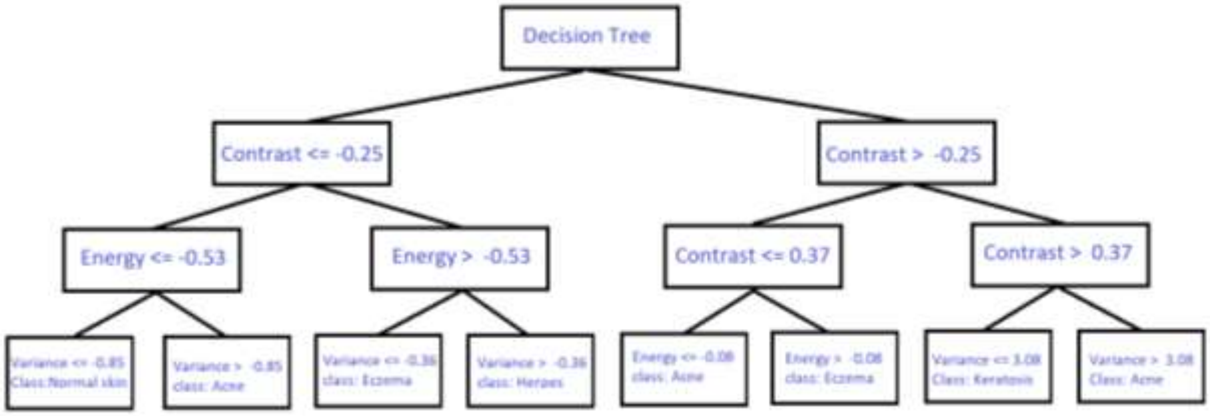

Decision Tree
Contrast <= -0.25
Contrast > -0.25
Energy <= -0.53
Energy > -0.53
Contrast <= 0.37
Contrast > 0.37
Variance <= -0.85 Class:Normal skin
Variance > -0.85 class: Acne
Variance <= -0.36 class: Eczema
Variance > -0.36 class: Herpes
Energy <= -0.08 class: Acne
Energy > -0.08 class: Eczema
Variance <= 3.08 Class: Keratosis
Variance > 3.08 Class: Acne

Fig 5.6: Classification Tree obtained by Decision Tree Algorithm

Fig 5.6 Represents a tree which determines the classification of disease based on respective image features ranges and based on the particular feature range disease is determined.

Table 5.4: Confusion matrix for SVM algorithm

	Class:1	**Class:2**	**Class:3**	**Class:4**	**Class:5**	**Classification overall**
Class:1	33	1	0	9	2	45
Class:2	0	30	17	0	0	47
Class:3	0	14	28	0	1	43
Class:4	9	0	0	32	0	41
Class:5	0	1	0	0	18	19
Truth overall	42	46	45	41	21	195

Table 5.4 indicates the SVM classification using linear kernel where data is linearly separable, diagonal elements of obtained matric indicates correct predictions which sum up to give 141 i.e., total 141 predictions were correct out of 195 and giving the total accuracy as 72.3%.

Accuracies obtained using different Machine Learning Algorithms are in the below table

Table 5.5: Accuracies of different ml Algorithms

Machine Learning Algorithm	**Accuracy Obtained**
Logistic Regression	71.79%
Support Vector Machine (SVM)	72.30%
Decision Tree	62.56%

From table 5.5, among all the machine learning Algorithms i.e., Logistic Regression, Support Vector Machine (SVM) and Decision Tree (DT) and highest accuracy was obtained by using Linear kernel in Support Vector Machine Algorithm with train size having 85% and test size having 15% of the given image extracted feature values.

CHAPTER 6
CONCLUSION

Machine learning algorithms were used along with image processing techniques for the detection of skin diseases. Here we used herpes, keratosis, eczema, and acne as the four diseases. In order to classify the type of disease using machine learning. these algorithms employ feature values from effected images as input. The choice of features is crucial for detecting skin diseases. In this project work, features such as entropy, variance, contrast, and energy are used to build machine learning algorithm such as logistic regression, Support Vector Machine, and Decision Tree. Accuracy is used to test the performance of the chosen algorithms.

Our findings indicate that the Support Vector Machine performed better when classifying two and three diseases, but as the number of diseases increases, its accuracy decreases. For two disease classifiers, logistic regression produced superior results. We used textural features to categorize skin diseases because the colors of various skin conditions are quite similar. With more datasets and richer characteristics, this work may be further enhanced. We can also develop some fundamental skin disease treatments.

REFERENCES

[1] K. V. Swamy, B. Divya, “Skin Disease Classification using Machine Learning Algorithms”, 2021 2nd International Conference Communication Computing and Industry 4.0 (C2I4), 2022, doi: 10.1109/C2I454156.2021.9689338 .

[2] V. B. Kumar, S. S. Kumar and V. Saboo, "Dermatological disease detection using image processing and machine learning", 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), doi: 10.1109/ICAIPR.2016.7585217

[3] P. R. Hegde, M. M. Shenoy and B. H. Shekar, "Comparison of Machine Learning Algorithms for Skin Disease Classification Using Colour and Texture Features", 2018 International Conference on Advances in Computing Communications and Informatics (ICACCI), doi: 10.1109/ICACCI.2018.8554512

[4] D.M.T. Rasanga, G.K.A.A. Tharushika, Ishara Weeruthunge, P. Bandara, “Machine Learning-Based Skin and Heart Disease Diagnose Mobile App”, 2022, doi: 10.1109/ECAI52376.2021.9515126

[5] N.V Kumar, P.V. Kumar, K. Promodh, y> Karuna. ”Classification of Skin Diseases using Image Processing and SVM”, 2019, DOI: 10.1109/ViTECoN.2019.8899449

[6] Honey Janoria, Jasmine Minj and Pooja Patre, "Classification of Skin Disease from Skin images using Transfer Learning Technique", 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), doi: 10.1109/ICECA49313.2020.9297567.

[7] T. Chauhan, S. Rawat, S. Malik and P. Singh, "Supervised and unsupervised machine learning based review on diabetes care", 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), doi: 10.1109/ICACCS51430.2021.9442021.

[8] Khushi Kumari Jha, Roshan Jha, Ankita Kumari Jha, Md Al Mahedi Hassan, Saurav Kumar Yadav, Tr Mahesh, "A Brief Comparison On Machine Learning Algorithms Based On Various Applications: A Comprehensive Survey", 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), doi: 10.1109/CSITSS54238.2021.9683524.

[9] P. Singh, S. P. Singh and D. S. Singh, "An introduction and review on machine learning applications in medicine and healthcare", 2019 IEEE Conference on Information and Communication Technology, doi: 10.1109/CICT48419.2019.9066250.

[10] Saibee Alam , Pankaj Mohrut,A Review on “Breast Cancer Stroma Maturity Classification Using DtG Filter and SVM”, 2019 International Conference on Intelligent Sustainable Systems (ICISS), doi: 10.1109/ISS1.2019.8908047.

[11] Srividhya E, Muthukumaravel A , Feature Extraction of Tongue Diseases Diagnosis Using SVM Classifier , 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), doi: 10.1109/ICCIKE47802.2019.9004326.

[12] Hansen F Charlie , Gloreine Dela Cruz , Testing Of Quality Of Water Using SVM, 2021 IEEE International Conference on Electronic Communications, Internet of Things and Big Data, doi: 10.1109/ICEIB53692.2021.9686418.

[13] Bhana Panwar, Gaurav Dhuriya, Prashant johri, Sudeept Singh Yadav, Nitin Gaur, Stock Market Prediction Using Linear Regression and SVM, 2021 International Conference on Advance Computing and Innovative Technologies in Engineering(ICACITE), doi: 10.1109/ICACITE51222.2021.9404733.

[14] Liu Lei, Prediction of Score of Diabetes Progression Index Based on Logistic Regression Algorithm, 2020 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), doi: 10.1109/ICVRIS51417.2020.00232.

[15] Velery Virgina Putri Wibowo, Zuherman Rustam, Afifah Rofi Laeli, Alva Andhika Sa'id, Logistic Regression and Logistic Regression-Genetic Algorithm for Classification of Liver Cancer Data, 2021 International Conference on Decision Aid Sciences and Application (DASA), doi: 10.1109/DASA53625.2021.9682242.

[16] Xiaonan Zou,Yong Hu,Zhewen Tian,Kaiyuan Shen, Logistic Regression Model Optimization and Case Analysis, 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), doi: 10.1109/ICCSNT47585.2019.8962457.

[17] Lei Liu, Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning, 2018 International Conference on Robots & Intelligent System (ICRIS), doi: 10.1109/ICRIS.2018.00049

[18] Tzuu-hseng S.Li,Huan-Jung Chiu,Ping-Huan Kuo, Hepatitis C Virus Detection Model by Using Random Forest, Logistic-Regression and ABC Algorithm, IEEE Access, doi: 10.1109/ACCESS.2022.3202295.

[19] Dianwei Chi, Research on the Application of K-Means Clustering Algorithm in Student Achievement - 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), doi: 10.1109/ICCECE51280.2021.9342164.

[20] Afroj Alam, Dr. Mohd Muqeem Integrated k-means Clustering with Nature Inspired Optimization Algorithm for the Prediction of Disease on High Dimensional Data - 2022 International Conference on Electronics and Renewable Systems (ICEARS), doi: 10.1109/ICEARS53579.2022.9752026.

[21] Pranab Sharma, Advanced Image Segmentation Technique using Improved K Means Clustering Algorithm with Pixel Potential. 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), doi: 10.1109/PDGC50313.2020.9315743.

[22] Md. Touhidul Islam, Sanjida Reza Rafa, Md. Golam Kibria, Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means, 2020 23rd International Conference on Computer and Information Technology (ICCIT), doi: 10.1109/ICCIT51783.2020.9392655.

[23] Harshitha Naidu Ravuvar, Haritha Goda, Sumathi R, P.Chinnasamy, Smart Health Predicting System Using K-Means Algorithm, 2020 International Conference on Computer Communication and Informatics (ICCCI -2020), doi: 10.1109/ICDMAI.2017.8073517.

[24] Phenyo Phemelo Moletsane, Tebogo Judith Motlhamme, Reza Malekian, Dijana Capeska Bogatmoska, A Review on Linear regression analysis of energy consumption data for smart homes,2018 41st International Convention on Information and

Communication Technology, Electronics and Microelectronics (MIPRO), doi: 10.23919/MIPRO.2018.8400075.

[25] Maria Susan Anggreainy, Hanif Musyaffa, Abdullah M. Illyasu, COVID-19 Patient Mortality Risk Classification Using Linear Regression and Exponential Smoothing Methods,2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), doi: 10.1109/ISMODE53584.2022.9742967.

[26] Nur Nafi'iyah, Kemal Farouq Mauladi, Linear Regression Analysis and SVR in Predicting Motor Vehicle Theft,2021 International Seminar on Application for Technology of Information and Communication (iSemantic), doi: 10.1109/iSemantic52711.2021.9573225.

[27] Asyraf Hakimi Abu Bakar, Najmuddin Mohd Hassan;Ammar Zakaria, Khairul Anwar Abdul Halim, Ahmad Ashraf Abdul Halim, Jaundice (Hyperbilirubinemia) detection and prediction system using color card technique,2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA), doi: 10.1109/CSPA.2017.8064952.

[28] Keren He, Cuiwei He, Housing Price Analysis Using Linear Regression and Logistic Regression: A Comprehensive Explanation Using Melbourne Real Estate Data, 2021 IEEE International Conference on Computing (ICOCO), doi: 10.1109/ICOCO53166.2021.9673533.

[29] N. Rochmawati, H.B. Hidayati, Y. Yamasari, W. Yustani, L. Rakhmawati, H.P.A. Tjahyaningtijas, Y. Anistyasari, “Covid Symptom Severity using Decision Tree”, 2020 3rd International Conference on Vocational Education and Electrical Engineering (ICVEE), 2020 , DOI: 10.1109/ICVEE50212.2020.9243246 .

[30] N. Ilyasova, N. Demin, Alexandr S., R. Paringer, “Fundus Image Segmentation using Decision Trees” , DOI: 10.1109/ITNT49337.2020.9253229.

[31] R, Euldji , M. Boumahdi, , M. Bachene , “Decision-making based on decision tree for ball bearing monitoring” , 2020 2nd Internantional Workshop on Human-Centric Smart Environment for Health and wealth being (IHSH) , OI: 10.1109/IHSH51661.2021.9378734.

[32] Huda. Kutrani, SariaEltahi , “Decision Tree Algorithms for Predictive Modeling in Breast cancer Treatment” , 2022 IEEE 2nd Internantional Maghreb meeting of the Conference on Science and Techniques of Automatic Control and Computer Engineering (MI-STA) Sabratha, Libya 23-25 May 2022, DOI: 10.1109/MI-STA54861.2022.9837762.

[33] A.Rajeshkanna , K. Arunesh , “ID3 Decision Tree Classification: An Algorithmic Perspective based on Error Rate”, 2020 , Proceedings of the Internantional Conference on Electronics and Sustainable Communication System (ICESC), DOI: 10.1109/ICESC48915.2020.9155578.