**SKIN CANCER PREDICTION USING MACHINE LEARNING MODELS**

Yashaswini Suresh

Vishwaradhya

***Abstract-* When skin cancer, a frequent kind of cancer, is found early, the chance of survival is increased. The objective is to develop subset of machine learning(ML) model that can classify images of dermal cells and detect skin cancer. Model-driven cloud architecture is used to create models that help to increase the accuracy of the skin cancer prediction. Subset of ML model techniques is crucial to this architecture. The project offers a demonstration of how to build models and use them to classify images of dermal cells. The Jupyter Notebook is used for this. The Kaggle website's dataset extraction tool was used. The quality and clarity of the membrane lesions were improved in the proposed work by eliminating artifacts, skin color, hair, and other components during pre-processing. The deep learning model created here displayed a 99.77% area of the curve when tested on common datasets. Using model-driven architecture, a practitioner may quickly produce subset of machine learning models for skin cancer prediction.**

***Keywords- Jupyter Notebook, Kaggle, Lesion, Melanoma***

I. INTRODUCTION:

The skin, or exterior layer of the body, is the biggest human body's organ. The skin's ectodermal tissues, which can have up to seven layers, protect the internal organs, muscles, bones, and ligaments beneath. Skin serves as a barrier between the outside world and the human body, aids in maintaining body temperature, and allows for the perception of touch, cold, and heat. Membrane lesions are theparts of the skin that are abnormal by difference to other parts of the skin. Infections that occur on or inside the skin are the fundamental and primary causes of membrane lesions. Membrane lesions can be divided into two categories: primary (which are present at birth or develop over the course of a lifetime) and secondary (which are brought on by improper handling of the primary membrane lesion) as shown in figure1. Skin cancer may result from either category, and more than two point nine million people receive a skin cancer diagnosis each year. In India, about 5000 skin cancer patients are hospitalised each year, and over 4000 people pass away from the disease. Three categories can be utilized to categorize skin tumors: Squamous cell cancer (SCC), melanoma, and basal cell carcinoma (BCC) When a tumor is diagnosed as malignant, a very deadly kind of skin cancer that spreads quickly to other parts of the body, it is regarded as cancer.An external file that holds a picture, illustration, etc.
Object name is ijerph-18-05479-g001.jpg

Fig.1 Skin Cancer Predictor steps in the extraction and implementation of data

In contrast, benign tumors are not as hazardous because they grow without spreading. Because the membrane lesion is examined with the naked eye and features cannot be precisely seen, maltreatment and eventually death happen from manual detection of skin cancer. Early precise skin cancer screening can boost survival rates. Automatic detection is hence more dependable for improved accuracy and effectiveness. Dermoscopy techniques are created to acquire the clear membrane lesion spot, and by reducing reflection, the visual effect is improved. However, it might be challenging to automatically identify membrane lesions because of artifacts, low contrast, skin tone, hairs, veins, and comparable visuals of melanoma and non-melanoma. All of this can be lowered with the aid of pre-processing methods. To find out the precise location of the membrane lesion, the pre-processed membrane lesion image is segmented. Multiple segmentation techniques exist, including the wavelet algorithm, basic global thresholding, region-based segmentation, watershed algorithm, snakes method, Otsu method, active contours, and geodesic active contours, among others. The process of segmentation makes use of geodesic active contour. SVM, KNN, Nave Bayes, and neural networks are only a some examples of the various types of classifiers. We applied characteristics to SVM, KNN, and Nave Bayes classifiers and obtained accuracy of 96%, 84%, and 76%, respectively.

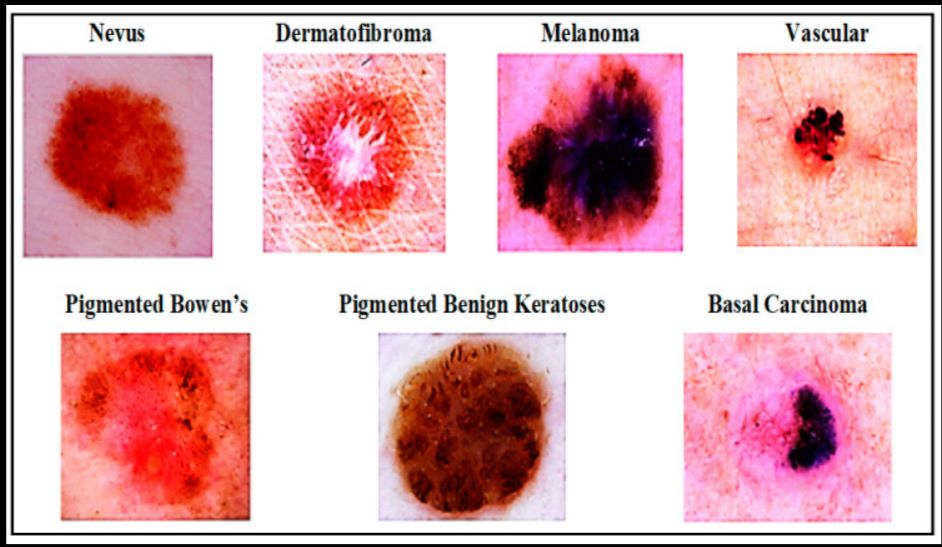
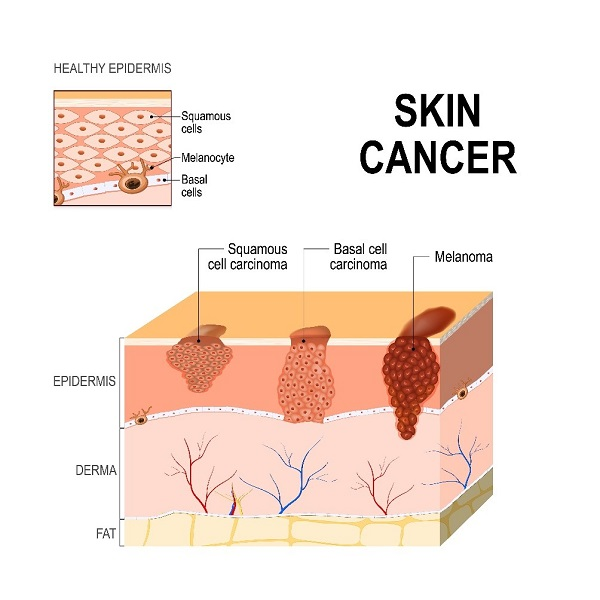


Fig.2 Types of skin cancer that can seem on the exterior of the skin

A significant and frequent condition is skin cancer. Over 5.4 million new cases of skin cancer are diagnosed once a year only in the USA. The figures are equally alarming on a global basis. Recent estimates claim that between 2008 and 2018, there was a 53% increase in the annual number of new melanoma cases. The likelihood that this illness will result in death is projected to rise over the next ten years. The survival percentage, if discovered later, is fewer than 14%. However, the survival probability is approximately 97% if skin cancer is discovered in its earliest stages. Early skin cancer detection is necessary as a result. This study resolves the issue of early diagnosis with greater accuracy.

It has been discovered that a qualified dermatologist typically follows a instruction of steps, starting with an unassisted eye visual evaluation of suspicious lesions, next dermoscopy (magnifying lesions microscopically), and lastly biopsy. The patient might go on to more advanced phases, and time would be lost. A proper diagnosis also depends on the clinician's skill. Even the best dermatologists can only accurately identify skin cancer in fewer than 80% of cases. In addition to these difficulties, the world's public healthcare systems are lacking in skilled dermatologists. It has taken a lot of work to develop computer image analysis codes and algorithms that can swiftly identify skin cancer in its early stages and solve few of the problems mentioned above.

The bulk of these computer solutions required correctly dispersed data since they were parametric. Because the nature of the data cannot be controlled, these procedures wouldn't be sufficient to accurately identify the conditionas shown in figure 2. On the other hand, non-parametric solutions don't rely on the need that the data have a normal distribution. Subset of machine learning is employed in this study to supplement the dermatologist's assistance. The method's fundamental premise is to train a computer to recognize the problem by looking at skin cancer photos. The presentation is innovative since the computer model may be developed without any programming knowledge. The average accuracy of this model's diagnoses is determined to be 98.89%, with 100% being the best.

Fig.3 Skin cancer depiction on a cross sectional view of the skin layers

The machine-assisted diagnosis presented here addresses the challenges of latency, accuracy, and a shortage of dermatologists in public health. Numerous research have been done on the subject of image categorization and skin cancer diagnosis, as be able to be publicized in figure 3. References give a comprehensive rundown of these methods. Although each of these studies used cutting-edge methods at the time, they all claimed greater performance. It is a widespread practice to categorize photos using Bayesian classifiers, decision tree algorithms, support vector machines, and other AI-based techniques. But one thing that all of these publications have in common is that they all claim to be written by authorities in the fields of software engineering and computers. A basic understanding of programming in languages like Java, R, and Python is required to build any of these diagnostic models. This article describes the methods for developing deep learning-based image classification models for detecting skin cancer without any prior programming knowledge. The major goals of this work are to:

1. Make it possible for researchers and practitioners to create subset of machine learning models using a straightforward plug-and-play method.

2. To use subset of machine learning to more precisely classify the cell images and identify cancer.

II. LITERATURE REVIEW

The correct disease diagnosis is among the most important steps in medical care. When it comes to diagnosis, dermatology among the most unpredictable and difficult specialties. Dermatologists frequently need more tests, a evaluation of the patient's medical history, and other information to guarantee a correct diagnosis. Finding a technique that can swiftly and accurately guarantee a reliable diagnosis is crucial. Over the years, a numerous methods have been created to make machine learning-based diagnosis easier. The produced systems, however, lack some qualities, such great precision. This paper suggests a MATLAB-based system that can recognise membrane lesions and categorise them as benign or normal[1].

The K-nearest neighbour (KNN) technique is used in the cataloging process to distinguish between healthy skin and malignant membrane lesions that suggest pathology. KNN is employed because it offers extremely accurate outcomes while being time-efficient. The system's classification accuracy for cutaneous lesions was 98%.

The most lethal and common kind of cancer is skin cancer. When found in its earliest stages, melanoma, the most serious kind of skin cancer, is completely curable. In order to diagnose melanoma, it is essential to find melanocytes in the epidermis. The segmentation approach employed in this piece of writing is the watershed segmentation method. Feature extraction is applied to the segments that were extracted[2]. Shape, the ABCD rule, and the GLCM were among the features that were extracted. The consequential features are then used to categorize the data. The classifiers include SVM (Support Vector Machine), Random Forest, and kNN (k Nearest Neighbor). The SVM classifier was demonstrated to provide superior classifications of membrane lesions when compared to other classifiers.

Skin cancer is a highly dangerous condition that poses a significant risk to people's lives. It occurs due to abnormal growth in melanocytic cells, resulting in malignant tumors known as melanoma. Genetic factors and exposure to UV radiation contribute to the development of melanoma, leading to the look of dark brown or black lesions on the skin. Early detection of melanoma is crucial for successful treatment and a favorable prognosis[3]. However, traditional diagnostic methods such as biopsies are invasive, uncomfortable, and time-consuming as they require laboratory analysis.

To address these challenges, the field of dermatology has turned to computer-aided diagnosis systems. Dermoscopy, a technique that captures detailed images of the skin, is employed in computer-aided diagnosis. In this research paper, the skin images undergo preprocessing to enhance their quality and remove noise. Subsequently, an image segmentation approach is utilized to separate the lesion area into its individual mechanism, allowing for the extraction of specific characteristics. These extracted features are then utilized in the identification of the skin image, distinguishing between normal skin and melanoma skin cancer, using advanced machine learning techniques for example support vector machines and K-nearest neighbor classifiers.

The findings from this proposed system demonstrate that comparing the results of support vector machines and K-nearest neighbor classifiers leads to the highest precision in skin cancer identification. This highlights the potential of computer-aided diagnosis in enhancing the correctness and efficiency of skin cancer detection, ultimately aiding in early intervention and improved patient outcomes.Top of Form

Skin cancer is the most prevalent and fatal type of cancer. Melanoma is the most dangerous sort of skin cancer and is totally curable when detected in its early stages. The discovery of melanocytes in the epidermis is a fundamental step in the identification of melanoma. The watershed segmentation method is used for segmentation in this paper. Feature extraction is useful to the extracted segments. Shape, ABCD rule, and GLCM are the structures that were retrieved. The features that were haul out are then used to classify data. The classifiers are SVM (Support Vector Machine), Random Forest, and kNN (k Nearest Neighbour). The SVM classifier proved to produce better results for the categorization of membrane lesions than other classifiers[4].

Skin cancer is a highly dangerous form of cancer primarily caused by unrepaired DNA damage in skin cells, leading to genetic mutations on the skin. Detecting skin cancer early is crucial due to its higher curability in the near the beginning stages and its potential to spread to other parts of the body. Early identification of skin cancer symptoms is necessary given the rising number of cases, high mortality rates, and expensive medical treatments. To embark upon these challenges, researchers have developed various techniques for untimely recognition of skin cancer. These techniques involve analyzing lesion characteristics such as symmetry, color, size, shape, and utilizing frameworks to differentiate between benign and malignant skin cancer. This study provides an in-depth exploration of subset of ML methods for early detection of skin cancer, examining research articles from reputable publications. The findings are presented using visual aids, graphs, tables, and methodologies to facilitate understanding [5].

Our skin shields our body from the sun's heat, light, and other dangers. Skin cancer is one of the diseases that can damage the skin. A mole with an uneven shape that is larger than a pencil pad may be the first sign of skin cancer. This study's main focus is on non-invasive methods for identifying and diagnosing skin cancer. The ABCD-Rule of Dermoscopy's asymmetry, border, and diameter criteria are used to regulate the geometric characteristics of moles suspected of having skin cancer. The characteristics loaded in the dataset for categorization include the biggest and smallest diameters, the irregularity index, and the equivalent diameter [6]. The k-Nearest Neighbors (k-NN) method is used to categorize images of moles. According to data, the total classification accuracy was 86.67%.

In today's rapidly expanding population, skin cancer affects people of all ages and not only the elderly. Dermatoscopic images are used to categorise a person's skin cancer into seven different categories. The HAM10000 (Human-Against-Machine with 10,000 training images) data collection is used to address this issue. The finalised dataset consists of 10001 dermatoscopic images, which are made publicly available through the ISIC repository as a ready set for academic machine learning applications. Dermatoscopic images are used to categorise a person's skin cancer into seven different categories.A person will learn from this research if they are at risk for mounting skin cancer of any kind, providing them with some peace of mind before they visit a doctor[7].

Skin cancer is both the worst and most prevalent type of cancer. Melanoma, the most dangerous form of skin cancer, is totally treatable when discovered in its earliest stages. Finding melanocytes in the epidermis is critical for the analysis of melanoma. This article uses the watershed segmentation method as its segmentation strategy. Feature extraction is applied to the extracted segments. Shape, ABCD rule, and GLCM were some of the features retrieved[8]. The data is then categorized based on the deduced qualities. kNN (k Nearest Neighbor), Random Forest, and SVM (Support Vector Machine) are the classifiers. The SVM classifier regularly produced classifications of membrane lesions that were superior to those produced by other classifiers.

Due to the typical standard substantial skin patches, it might be difficult to identify and monitor benign moles and skin malignancies. Actually, there isn't much variation in the appearance of membrane lesions, and there isn't much information available. Basal Cell Carcinoma (BCC), Melanoma, and Squamous Cell Carcinoma (SCC) are three of the seven basic kinds of skin cancer, but Melanoma is the most dangerous and has a low survival rate.

The photos used in this training are trained from start to finish using a convolution neural network to identify membrane lesions. In direction to train a convolution neural network (CNN), 10,000 clinical images were used as the dataset. Resources and Techniques Image pre-processing, which categorizes photos and removes duplicates, and sharpening, which resizes the skin image, are often the two key processes in the process of finding skin cancer[9]. The research that follows describes a technique for utilizing subset of machine learning to distinguish between high-level membrane lesions and determine malignancy: 1) Development of a neural network that can precisely identify the edge of a significant lesion; 2) Development of a model that can be applied to mobile devices. The transfer learning approach, which is based on deep neural networks and fine turning, helps to achieve high prediction accuracy.

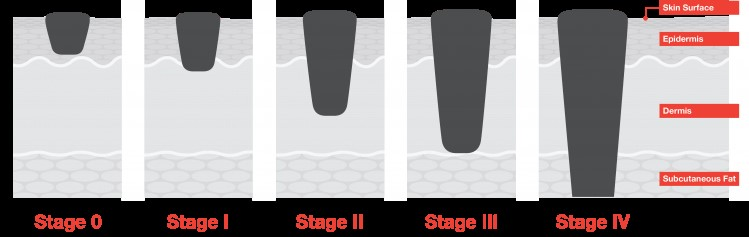
Results: The dataset consists of 10,000 photos that are divided into two folders. A data frame contains the data's descriptive information. 374 melanoma photos, 254 seborrheic keratosis images, and 1372 nevus images are included in the total of 10,000 dermoscopic images. Top-2 accuracy and Top-3 accuracy have been calculated using transfer learning validation loss. The outcome has been contrasted with several models. Conclusions: Eczema, acne, malignant, and benign membrane lesions can all be categorised using the suggested system. The suggested research examines the characteristics that the deep convolutional neural network has learned. The datasets were split into seven distinct categories once the attributes were retrieved. The data was validated and trained based on such categories.

Skin cancer is an overgrowth of skin tissue that affects the skin. It has an uneven structure with cell differentiation at different levels in the chromatin, nucleus, and cytoplasm. It is also expansive, infiltrative, and causes damage to the surrounding tissue. Skin cancer also spreads through blood arteries and lymph vessels. The biopsy method of skin cancer diagnosis is regarded as less effective because it is expensive and potentially harmful to human skin samples. For so, we want an efficient and precise system for classifying different forms of skin cancer. The health industry has made extensive use of machine learning. Random Forest is individual of the machine learning techniques. The histogram colour feature extraction, hue moment shape extraction, and haralick texture extraction will all be done in this study. The classification of the image will also application of the Random Forest technique [10]. The best accurateness significance was obtained using the approach of collecting histogram features and categorizing the data by means of Random Forest, which was 0.850822. A unique part of this research is the use of more varied feature extraction algorithms and the increased accuracy results in comparison to earlier studies. Future research is predicted to employ subset of machine learning algorithms using CNN (Convolutional Neural Network) architecture in order to enhance accuracy results and include application designs for the application of models created in the study so they may be used immediately by the medical team.

The objective of this project is to develop a case-based method for diagnosing skin cancer using user data. By using conversational case-based reasoning, users are assisted in expressing their problem through the question-dialog process. DePicT is a knowledge-based method for identifying and forecasting illnesses that combines text information from patient health records with image classification[11]. It uses both textual and visual data sources, and then uses case-based reasoning to give a recommendation. The Case-based Melanom (CBMelanom) method is employed in this article to communicate with users, record their skin-related issues in a conversation (question-dialog), and suggest the most appropriate course of action.

Skin malignant growth is one of the deadliest diseases, and it kills more people than it used to since people are less aware of the symptoms and don't take precautionary measures. In order to stop the spread of cancer, early identification at an early stage is essential as in figure 4. Other types of skin cancer with several classifications include melanoma, nevi, and seborrheic keratosis[12]. The many types of skin cancer are identified and categorized in this study using artificial intelligence (AI) and image processing techniques. Dermoscopic pictures are considered as contributions during the pre-handling operations. the development of a convolutional neural network-based melanoma detection model. It is the primary cause of 75% of skin cancer deaths.

Using visual examination and manual analysis of pictures of membrane lesions, it has always been challenging to detect skin cancer. Examining membrane lesions by hand to check for melanoma may require a lot of time and effort. One of the most deadly diseases is skin cancer, particularly melanoma. Due to their similarities in color skin imaging, such as between carcinoma and nevi, it is more difficult to recognize and diagnose various membrane lesions[13]. As a result of technological advancement and the quick increase in processing power, numerous subset of machine learning models and machine learning techniques have developed for the interpretation of medical images, notably the images of membrane lesions. Early detection and sorting of membrane cancer allow for proper diagnosis and treatment for patients.



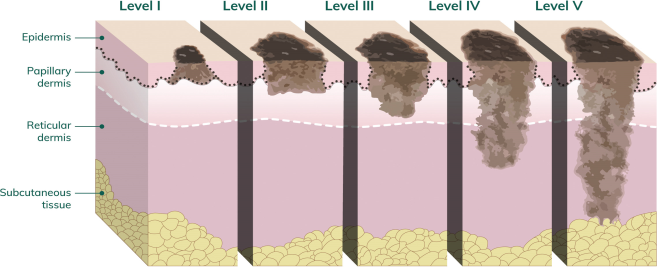


Fig.4 Different stages of skin cancer

For the most successful treatment of the two most common malignancies, skin and oral, early diagnosis is crucial. Using computer-aided cancer detection (CAD) and medical images, deep learning techniques have shown promising outcomes in detection of malignant malignancies. This work suggests a deep learning-based strategy for exploiting medical imaging to detect skin and mouth cancer. We explore numerous Convolutional Neural Network (CNN) models, including Graph Neural Network (GNN), ResNet, DenseNet, Inception, AlexNet, VGGNet, and Inception[14]. To enhance the quality and reduce noise, skin cancer and oral cancer photos are subjected to figure dispensation techniques such image scaling and image filtering.

III. METHODOLOGY

1. KNN CLASSSIFIER



Fig.5 KNN Classifier Representation as a block diagram

One of the simplest machine learning algorithms, K-Nearest Neighbor uses the supervised learning method.

The K-NN algorithm places the new case in the category that is mainly similar to the available categories based on the hypothesis that the new instance and the data are comparable to the examples that are already accessible.

The K-NN algorithm saves all the data that is available and categorizes new input based on similarity. This means that as fresh data is generated, it may be quickly categorized into a suitable category using the K-NN method as shown in figure 5.

Although the K-NN technique can be used for both classification and regression problems, it is more frequently utilized for classification issues. The K-NN algorithm is non-parametric, which means it doesn't make any assumptions about the underlying data as shown in figure 6.

As a consequence of saving the training dataset rather than instantly learning from it, the method is sometimes recommended to as a lazy learner. Instead, it performs an action while categorizing data by using the dataset.

The KNN method simply saves the information during the training phase, and when it receives new data, it categorizes it into a category that is quite similar to the new data as shown in figure 7.

Example: Let's say we have a picture of a species that resembles both cats and dogs, but we aren't sure if it is one or the other. Therefore, since the KNN algorithm is based on a similarity metric, we can utilize it for this identification. Our KNN model will guise for features in the new information set that are comparable to those in the photographs of cats and dogs, and based on those facial appearances; it will organize the data as belonging to either the cat or dog group.

**Data Aquisition**

**Data Processing**

**Feature Extraction**

* **Age**
* **Sex**
* **Trestbps**
* **Chol**
* **Baseline**
* **trt**

**KNN Classification for the data**

Fig.6. KNN Classifier Flowchart



Fig.7 Graphical understanding of KNN Classifier

2. DECISION TREE

Guided learning techniques called decision trees are recurrently used to address classification problems. They may also be used to address regression problems. It is a tree-structured classifier, with leaf nodes designating the classification outcome and interior nodes designating dataset characteristics, decision rules, and branching, respectively.

Decision Tree Terminologies(shown in figure 8)

* Root Node: The starting point of a decision tree, representing the entire dataset, which is then divided into homogeneous subsets.
* Leaf Node: The final output nodes of a decision tree that cannot be further divided.
* Splitting: The process of dividing the decision/root node into sub-nodes based on specific criteria.
* Branch/Sub Tree: A tree formed by cutting off a portion of the original tree.
* Pruning: The act of removing undesirable branches from a decision tree.
* Parent/Child Node: In a tree structure, the root node is well thought-out the parent node and all extra nodes are its child nodes.

In decision trees, the root node begins the tree and represents the entire dataset. Splitting occurs to create sub-nodes based on certain criteria. The leaf nodes are the final outputs and cannot be further divided. Branches or sub-trees are created by cutting off portions of the original tree. Pruning involves removing unwanted branches. The majority class for every node is represented by the node's color.

Based on the distinctiveness of the available dataset, judgments or tests are run as shown in figure 9.

The most important thing to keep in mind while developing a machine learning model is to select the optimal method for the dataset and task at hand. This also increases the generalization of the model and allows it to predict the new data with more accuracy.

**

Fig.8 Decision Tree Representation as a block diagram

The two rationales for employing the decision tree are as follows:

* Decision trees are typically designed to resemble how people think when making decisions, making them simple to comprehend.
* The decision tree's reasoning is clear because it has a tree-like structure.

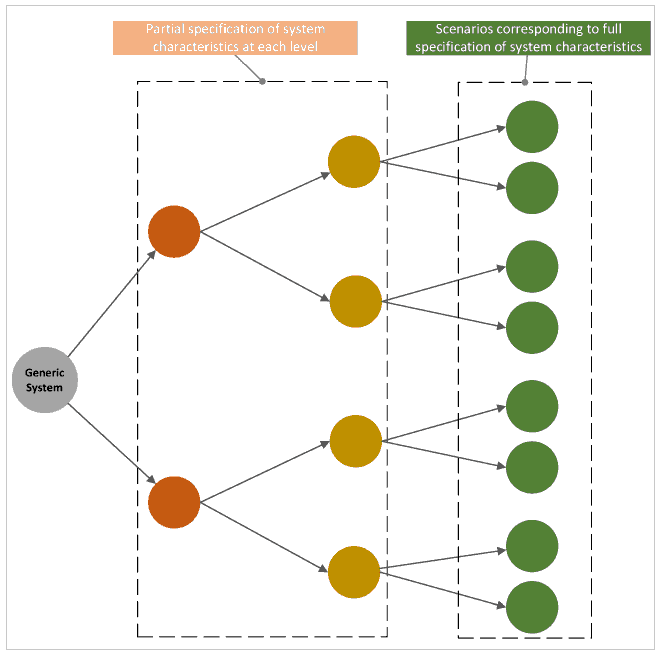


Fig .9 Understanding of Decision Tree Algorithm

The methodologies include the algorithm used, dataset used and flowchart of the data used and implemented. Below is the provided step by step explanation of the algorithm used. Algorithm used: The decision tree algorithm is an extensively used supervised learning technique employed for both classification and regression. It constructs a structured model resembling a flowchart, driven by input features.

1. Tree Construction: The algorithm commences by considering the full dataset as the source node, and selects the optimal feature for partitioning.

2. Feature Split: The chosen feature is utilized to split the data into subsets, thereby creating branches or paths within the decision tree.

3. Recursive Splitting: The process of feature splitting is iteratively applied to each subset until a predefined stopping criterion is satisfied.

4. Leaf Node Assignment: Leaf nodes are assigned class labels or regression values based on the majority class or mean assessment of the target variable within each respective subset.

5. Prediction: To make predictions, the algorithm traverses the decision tree by evaluating feature values and ultimately reaching a leaf node to achieve the final prediction as shown in figure 10.

Advantages: Easy to comprehend and interpret Accommodates numerical and categorical data Handles missing values gracefully Captures non-linear relationships effectively

Limitations: Prone to over fitting, necessitating proper regularization techniques - Can be sensitive to changes in the dataset, leading to instability. Exhibits partiality towards features with elevated cardinality or many levels in conclusion, decision trees offer versatility and transparency in model interpretation. However, caution must be exercised to address overfitting issues and effectively manage the algorithm's limitations.

1. Import necessary libraries and module: from sklearn. tree import Decision tree classifier, import pandas as pd and few other libraries are imported as shown in figure 11.

2. Read the dataset: use necessary commands to read the dataset provided and collect the information from it.

3. Preprocess the data: here the data collected is pre-processed so that feature extraction can take place easily.

4. Load the data: The data pre-processed now will be loaded into the training model to train the model properly.

5. Splitting the data: The data is splitted into training and testing data in proportion of 80:20 out of 100.

6. It checks for the target: if present then goes to next step or else declares it as unsupervised model as it does not have specified output. 7. Checks for target data: we have used the decision tree classifier therefore uses discrete data.

8. Training model: after all the necessary adjustments in the code we train the model using decision tree algorithm.

.

**Data Aquisition**

**Data Processing**

**Feature Extraction**

* **Age**
* **Sex**
* **Trestbps**
* **Chol**
* **Baseline**
* **trt**

**Decision Tree Classification for the data**

Fig.10 Decision Tree Classifier Flowchart (up)

9. Testing data: after training the model is tested and accuracy is obtained with necessary graphs. A small comparison plot is also included in algorithm.

10. Results: as last step the results are obtained and the model execution ends. Few additional steps of regularization imputation and many other commands are included in the code to improve the accuracy and prevent the model from overfitting and underfitting.

Import the required libraries(like pandas, matplotlib, etc.) to the jupyter notebook

Initialize the train and test datasets

Check the test data for errors and verify

Check the dataset Evaluate accuracy

Import the dataset by using its path Define Feature and Target

Plot the output and other required graphs using matplotlib

Fit the Data.Train and test the data

Data is improper

Verify the data again

Data is in its proper order

Fig.11 Flowchart of the designing process of the ML model

DATASET

The dataset currently used in the model training is taken from the Kaggle website as shown in figure12. Kaggle is a popular platform for data scientists and ML practitioners to discover and share datasets, as well as participate in data science competitions. It hosts a vast collection of datasets from various domains, allowing users to access and analyse realworld data. Kaggle datasets are typically provided in structured formats such as CSV, Excel, or SQL, and may contain a wide range of variables or features. These datasets cover diverse topics including healthcare, finance, social sciences, computer vision, natural verbal communication dealing out, and more. Users can explore and download datasets from Kaggle for their own analysis, model training, or research purposes. They can also contribute by uploading and sharing their own datasets with the Kaggle community. Kaggle datasets are accompanied by detailed descriptions, documentation, and often include pre-split train/test datasets to facilitate model development and evaluation. Additionally, many datasets come with sample code notebooks, known as kernels, which provide examples and insights on how to work with the data. By leveraging the vast collection of Kaggle datasets, data scientists can gain access to high-quality, real-world data to explore, analyse, and build predictive models or develop insights for a ample collection of applications. The dataset contains basic clinical features of skin cancer which includes sex, age, target, cholesterol, trestbps, localization, baseline and time. These features are utilized in the model training to predict the disease accurately.

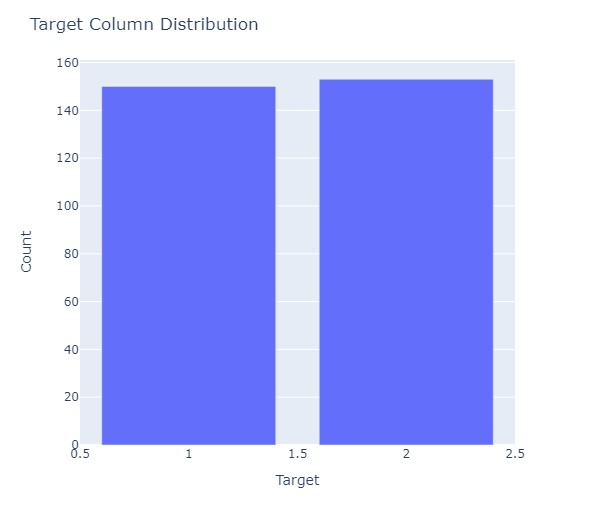


Fig.12 Dataset HAM10000 downloaded from Kaggle

V. RESULTS AND DISCUSSION

First, an interactive bar plot with the target\_counts DataFrame as the data source was created using plotly.express (px.bar()). The bars depict the distribution of the target column, while the x-axis depicts the unique target values and the y-axis the count of each target value. adds a title, x-axis label, and y-axis label to the plot's layout. utilizes the show() function to display the plot.

The below graph shown in figure 13, plots the training points from X\_train on a scatter plot, where the x and y coordinates come from the initial two features (X\_train[:, 0] and X\_train[:, 1]), and the colors come from a colormap specified by colors, which represents the standards of the third feature (X\_train[:, 2]). It uses plt.contourf() to draw the decision boundaries on the scatter plot, where the decision borders are specified by Z. The intelligibility of the filled sections is controlled by the alpha parameter.

****Fig. 13 Bar graph of count vs. target

It also sets the plot's title, y-axis label, and x-axis label. Plot is shown using plt.show().

A scatter plot containing training points and decision bounds is the code's output as shown in figure 14.

A scatter plot containing training points and decision bounds is the code's output. In the feature space that is defined by the initial and succeeding features, the scatter plot displays the allocation of the training points. The standards of the third characteristic (sex) are represented by the colors of the points. The projected class labels for the meshgrid points serve as the decision boundaries, which divide the feature space into several regions. The figure makes it easier to see how the decision tree classifier divides the many classes in feature space

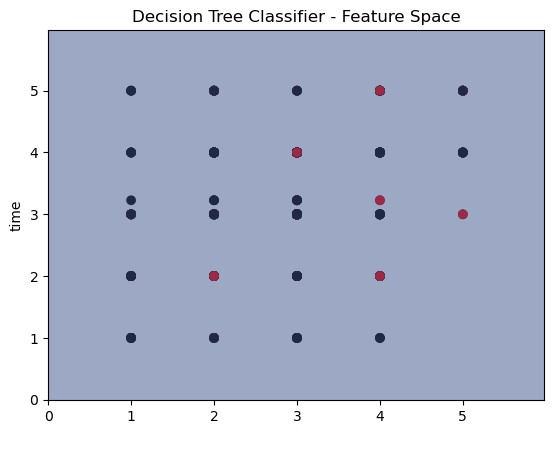


Fig.14 Feature space of decision tree classifier

The decision tree is shown visually as the code's output. The arrows show the flow from one node to another based on the decisions, and each node in the tree reflects a decision based on a feature as shown in figure 14.



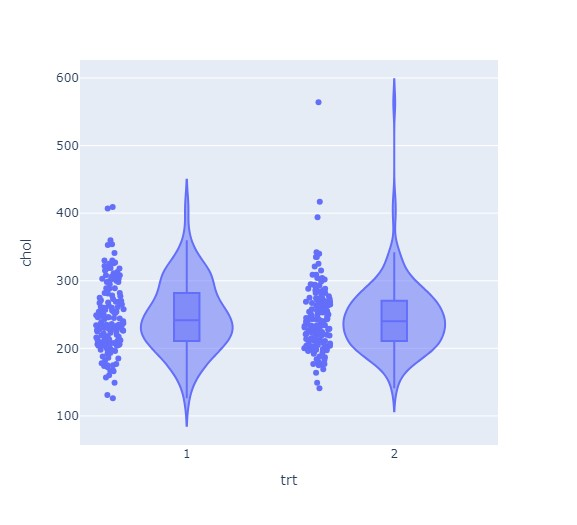
****

Fig.15 Logarithmic Trends of the features

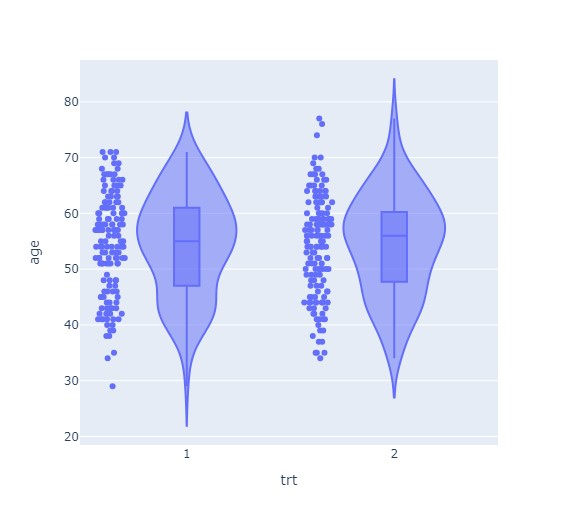
****

Fig.16 Violin plot of age feature

Fig.17 Violin plot of chol feature

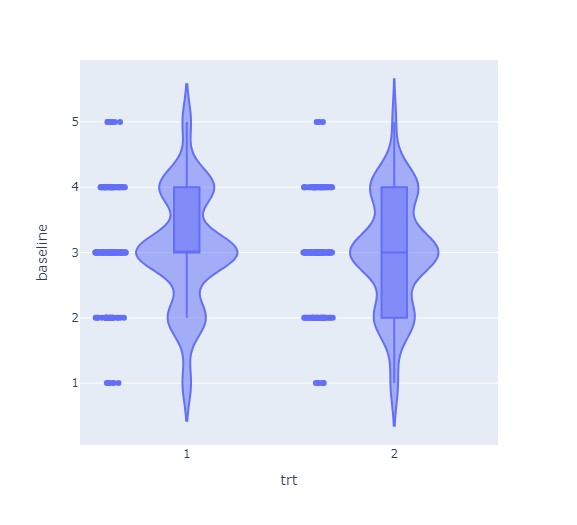
****

Fig.18 Violin plot of baseline feature

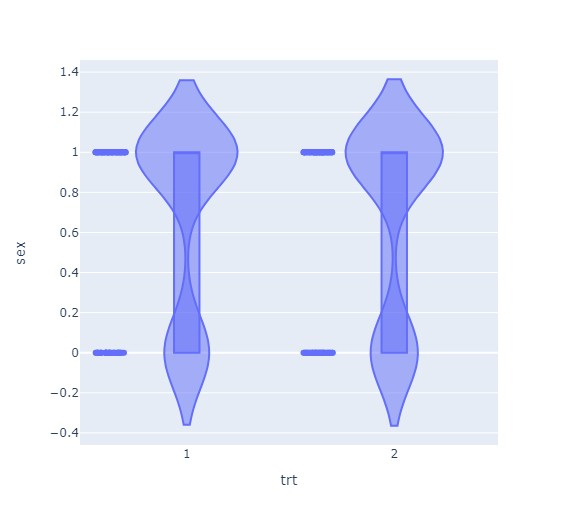
****

Fig.19 Violin plot of sex feature

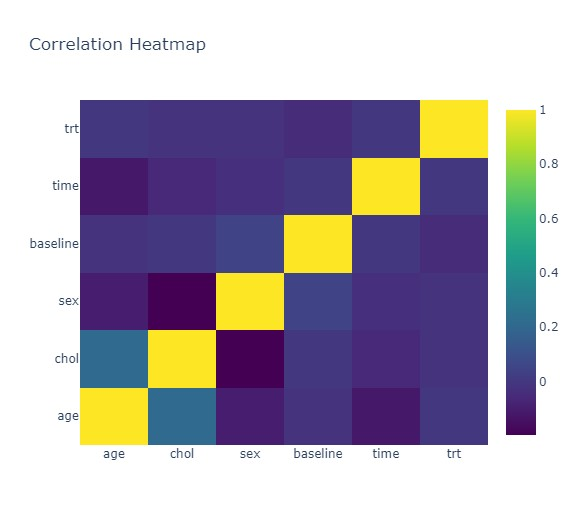


Fig.21 Correlation Heatmap in decision tree

The output of the code includes the violin plots(as shown in figures 16, 17, 18, 19 and 20), correlation heatmap, accuracy score, confusion matrix, and dataset table, providing insights into the data distribution, correlations, model performance, and dataset overview. This helps in better perceptive of the created machine learning model. The decision-making process of the decision tree classifier based on the characteristics and their significance in forecasting is depicted visually in the plot.

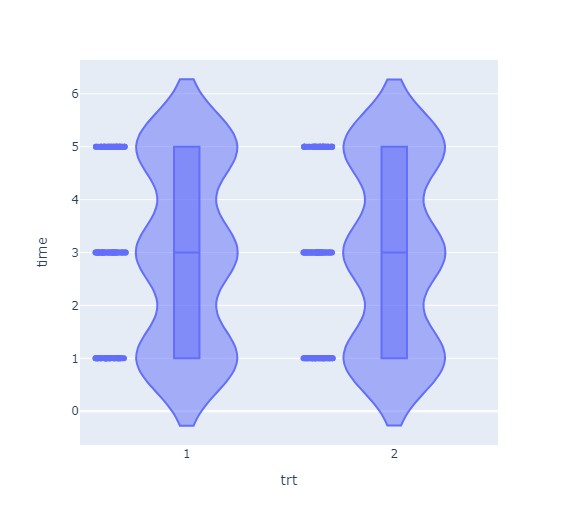


Fig.20 Violin plot of time feature

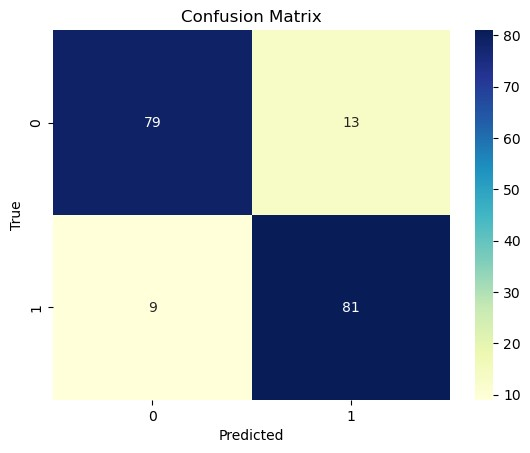


Fig.22 Confusion matrix with time against predicted

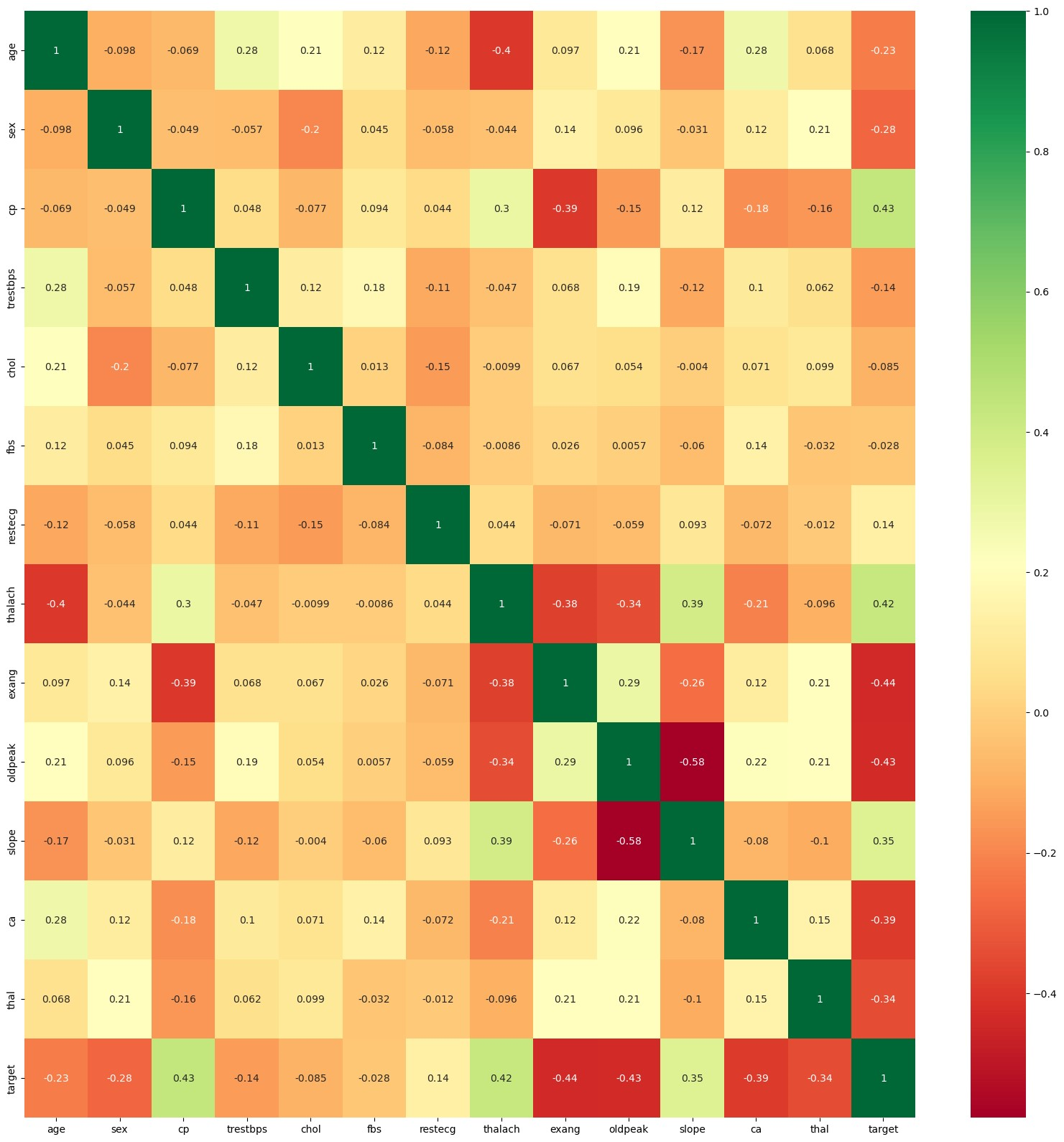


Fig.23 Correlation heatmap of all the features

A heatmap is a way to represent the viewer's behavioral data in the form of hot and cold bits using a color scheme from hot to coldas shown in figure 21. Warm colors represent areas with the most visitor interaction, red represent areas with the most interaction, and cool colors represent areas with the least interaction as shown in figure 23.

A table called a confusion matrix is used to describe how well a classification system performs. The output of a classification algorithm is shown and summarized in a confusion matrix. Figure 22 displays a confusion matrix where benign tissue is referred to be healthy and malignant tissue is seen as cancerous.

It uses sns.heatmap() to produce a heatmap. The correlation matrix of the chosen characteristics is used as the input data. The annot=True option adds the correlations' numerical values in each cell.

uses plt.show() to display the heatmap. The code's result is a heatmap that showsthe relationships between the various dataset characteristics The code's result is a heatmap that shows the relationships between the various dataset characteristics. The coefficient of correlation between two characteristics

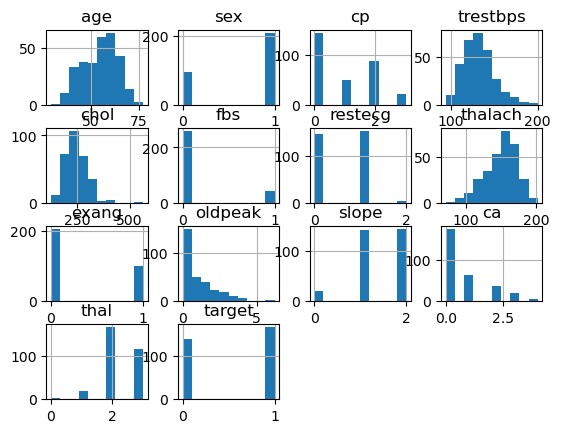


Fig.24 Individual histogram of all features

is shown in each heatmap cell. Each cell's color denotes the degree and direction of the link. Shades of green are used to symbolize positive correlations, whereas hues of red are used to depict negative correlations. The precise correlation coefficient is given by the numerical values in each

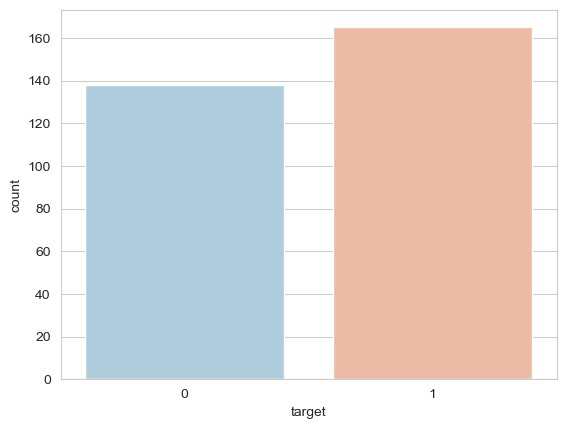
The above graph gives a histogram(as shown in figure 24) charted view of each feature in the knn classifier

Fig.25 Distribution graph of target vs. count

The output of the code is a countplot that displays the distribution of the 'target' column. Each bar represents the count of occurrences for each unique value in the 'target' column as shown in figure 25. The countplot provides insights into the class distribution and can be used to analyze the balance or imbalance of classes in the dataset.

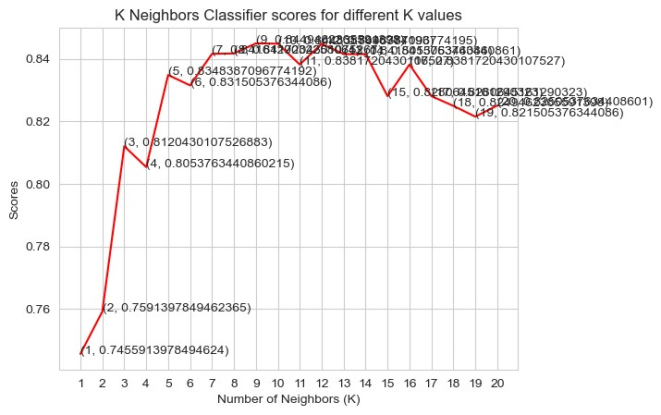


Fig.26 Line Plot of KNN Scores

The output of the code is a line plot that shows the scores of the K Neighbors Classifier for different values of K. The x-axis signifies the number of neighbors (K), and the y-axis represents the corresponding scores as shown in figure 26. Each point on the line represents a specific value of K and its corresponding score. The text labels provide additional information about the exact values of K and the corresponding scores at each poin

The code produces a graphic with six vertically stacked subplots, each of which shows the trend of a different characteristic over the dataset's index. The index is represented by the x-axis, while the feature values are shown by the y-axis.

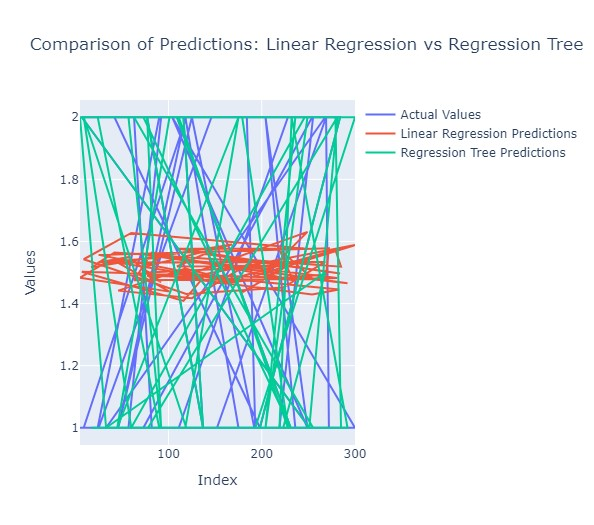


Fig.27 Interactive line graph of actual values(blue)

And prediction values(red and green)

The lines in the graphic show how the feature values varied and changed across the dataset in various ways.

Table I Comparison between knn and decision tree algorithms

|  |  |  |
| --- | --- | --- |
| Serial No. | Knn classifier | Decistion Tree Classifier |
| 1. | There is no real 'training' time since no computation is performed while training. All that needs to be done is storing the training data. | Training involves iteratively building a tree (considering the ID-3 algorithm) so considerably more time than k-NN. |
| 2. | High testing time since the distance needs to be computed between the test point and every training point. | Low testing time since it's a traversal down a tree. |
| 3. | The user needs to choose a distance metric in order to use a k-NN for testing. | No need to choose a distance metric since the splits will occur based on values inherent to each feature. |
| 4. | Updating an existing model with new data simply means adding that point to the existing dataset | An existing tree cannot be updated - an entire new tree needs to be created |
| 5. | Need to store all the training data in order to classify a new incoming point | No need to store the training data - only store the tree model |
| 6. | Accuracy on Training set was found to be 0.865 | Accuracy on Training set was found to be 0.845 |
| 7. | Accuracy on Test set was found to be 0.79 | Accuracy on Training set was found to be 0.7465 |
| 8. | Mean score was 0.832. | Mean score was 0.792. |

The output of the code is an interactive line graph that compares the actual values, predictions from Linear Regression, and predictions from Regression Tree models. It allows visual assessment of how well the models perform in predicting the target variable as shown in figure 27.Additionally, the mean squared errors for both models are printed, providing a quantitative measure of the prediction accuracy as shown in Table I.

VI. CONCLUSION AND FUTURE SCOPE

Diverse neural network algorithms for skin cancer recognition and organization have been covered in this systematic review research. These methods are all non-invasive. The method of detecting skin cancer involves numerous steps, together with preprocessing, picture segmentation, feature extraction, and classification. The categorization of lesion pictures using ANNs, CNNs, KNNs, and RBFNs was the main emphasis of this review. Each algorithm has benefits and drawbacks. The key to getting the best results is choosing the categorization method correctly. However, because it is more directly tied to computer vision than other neural networks, CNN performs better than other types of neural networks when categorizing picture data.

The majority of skin cancer detection research focuses on determining if a particular lesion picture is malignant. However, the most recent study is unable to respond to a patient's question about whether a certain skin cancer symptom manifests itself elsewhere on their body. The study has consequently far been confined to the particular issue of classifying the signal picture. Full-body photography may be used in future studies to help find the solution to this topic. The picture capture process will be automated and accelerated via autonomous full-body photography.

Subset of machine learning has lately given rise to the concept of auto-organization. The method of unsupervised learning known as "auto-organization" seeks to recognize characteristics and find relationships or patterns in the dataset's picture samples. Auto-organization approaches improve the level of features representation that is recovered by expert systems under the guise of convolutional neural networks [47]. Auto-organization is still a paradigm that is being researched and developed at this time. Even in the field of medical imaging, where the tiniest details of characteristics are vital for the accurate identification of disease, its research, however, can increase the accuracy of image processing systems in the future.

VII. ABBREVIATIONS

SVM - Support Vector Machine

ML - Machine Learning

KNN - K-Nearest Neighbors Algorithm

CNN - Convolutional Neural Network  
ANN - Artificial Neural Network  
RBFN - Random Block File Manager

ABCD - Asymmetrical, Border, Color, Diameter

GLCM - Gray level Co-Occurrence Matrix

REFERENCES

1. H. Kibriya, I. Abdullah and F. Kousar, "Melanoma Lesion Segmentation and Classification Using SegNet," 2023 4th International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 2023, pp. 1-6, doi:

10.1109/ICACS55311.2023.10089675.

1. H. T. Lau and A. Al-Jumaily, "Automatically Early Detection of Skin Cancer: Study Based on Nueral Netwok Classification," 2009 International Conference of Soft Computing and Pattern Recognition, Malacca, Malaysia, 2009, pp. 375-380, doi:
2. Dildar, M.; Akram, S.; Irfan, M.; Khan, H.U.; Ramzan, M.; Mahmood, A.R.; Alsaiari, S.A.; Saeed, A.H.M.; Alraddadi, M.O.; Mahnashi, M.H. Skin Cancer Detection: A Review Using Subset of machine learning Techniques. Int. J. Environ. Res. Public Health **2021**, 18, 5479.
3. A. Murugan, S. Anu H Nair, A. Angelin Peace Preethi, K. P. Sanal Kumar, 10.1109/SoCPaR.2009.80.Melanoma Skin Cancer Detection Using Knn And Svm Classifier
4. M. Vidya and M. V. Karki, "Skin Cancer Detection using Machine Learning Techniques," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2020, pp. 1-5, doi: 10.1109/CONECCT50063.2020.9198489.
5. M. Krishna Monika, N. Arun Vignesh, Ch. Usha Kumari, M.N.V.S.S. Kumar, E. Laxmi Lydia,Skin cancer detection and classification using machine learning,Materials Today: Proceedings,Volume 33, Part 7,2020,Pages 4266-4270
6. M. Ramachandro, T. Daniya and B. Saritha, "Skin Cancer Detection Using Machine Learning Algorithms," 2021 Innovations in Power and Advanced Computing Technologies (i-PACT), Kuala Lumpur, Malaysia, 2021, pp. 1-7, doi: 10.1109/i-PACT52855.2021.9696874.
7. Md. Hasan, Surajit Das Barman, Samia Islam, and Ahmed Wasif Reza. 2019. Skin Cancer Detection Using CNN. In Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence (ICCAI '19). Association for Computing Machinery, New York, NY, USA, 254–258.
8. Praveen Banasode et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. **1065** 012039 S. R and V. K, "Skin Cancer prediction using Machine Learning," 2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC), Bengaluru, India, 2022, pp. 1112-1115, doi: 10.1109/IIHC55949.2022.10060544.
9. N. Tyagi, L. Dhavamani, M. S. A. Ansari, B. Pant, D. K. J. B. Saini and J. A. Dhanraj, "Skin Cancer Prediction using Machine Learning and Neural Networks," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 271-275, doi: 10.1109/IC3I56241.2022.10073141.
10. B. Saju, V. Asha, S. C. Murali, V. D, V. Kumar and B. Nithya, "ML based Prototype for Skin Cancer Detection," 2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/C2I456876.2022.10051378.
11. S. Sharma, K. Guleria, S. Kumar and S. Tiwari, "Benign and Malignant Membrane lesion Detection from Melanoma Skin Cancer Images," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-6, doi: 10.1109/ICONAT57137.2023.10080355.
12. H. Kibriya, I. Abdullah and F. Kousar, "Melanoma Lesion Segmentation and Classification Using SegNet," 2023 4th International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 2023, pp. 1-6, doi:

10.1109/ICACS55311.2023.10089675.

1. A. Murugan, S. Anu H Nair, A. Angelin Peace Preethi, K. P. Sanal Kumar, 10.1109/SoCPaR.2009.80.Melanoma Skin Cancer Detection Using Knn And Svm Classifier