International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Skin Cancer Detection Using Machine Learning

Maridu.Bhargavi[1], Assistant Professor, Department of CSE,

Vignan's Foundation for Science, Technology & Research, Vadlamudi, Guntur Dt.,Andhra Pradesh.

V.Ujwala[2], **M. Hema Sree**[3], **Krishna priya**[4]

[2,3,4] UG Students, Department of CSE,
Vignan's Foundation for Science, Technology & Research, Vadlamudi, Guntur Dt.,Andhra Pradesh

ABSTRACT

skin cancer is a normal form of cancer that affects millions of people worldwide. Early diagnosis and detection are crucial in the successful treatment of skin cancer. Machine learning algorithms have been utilized in recent years to assist in the detection and diagnosis of skin cancer. This abstract presents a unique approach to skin cancer detection using machine learning. The suggested approach combines machine learning algorithms with image processing methods to analyse images of skin lesions and classify them among 9 different classes. The system employs an innovative feature selection method to identify the most significant features that aid in the classification process. These features include Edge detection, histogram, Texture analysis. The classification of images is done by using Decision Tree, KNN, Random Forest and Voting classifier algorithms. The proposed system was trained and tested on a dataset of skin lesion images obtained from various sources, including dermatology clinics and online databases. The findings show that the suggested strategy is highly accurate at identifying and categorising skin lesions. The unique strategy described in this abstract may increase the precision and effectiveness of skin cancer detection, resulting in earlier diagnoses and more effective therapies.

Keywords: Decision Tree, KNN, Voting classifier, Random Forest, GLCM

Introduction

The escalating prevalence of skin cancer over recent decades highlights the urgent need for effective preventive measures and treatments. As the most prevalent form of cancer worldwide, skin cancer affects approximately one in five Americans during their lifetime. This project aims to provide a comprehensive overview encompassing the causes, risk factors, diagnosis, and treatment options associated with skin cancer. Additionally, we will explore current advancements in skin cancer research and discuss potential directions for future studies. Skin cancer arises when abnormal cell growth occurs within the skin's cells. Melanoma, squamous cell carcinoma, and basal cell carcinoma are the three main kinds of skin cancer. Basal cell carcinoma, accounting for around 80% of cases, emerges from the basal cells located in the skin's bottom layer, known as the epidermis. Squamous cell carcinoma, comprising approximately 16% of cases, originates from the squamous cells present in the epidermis' upper layer. Although melanoma is the least common form of skin cancer, representing only 4% of cases, it is the most dangerous due to its potential to metastasize. Melanoma develops in the melanocytes, the pigment-producing cells in the skin. The primary cause of skin cancer is exposure to ultraviolet (UV) light from the sun or tanning beds. UV light damages the DNA of skin cells, resulting in genetic alterations and the eventual development of cancer. Additional skin cancer risk factors include having fair skin, a history of sunburns, a weakened immune system, and a family history of the illness. An assessment of the skin's physical condition is usually followed by a biopsy to determine whether skin cancer cells are present. Treatment options for skin cancer encompass surgical procedures, radiation therapy, and chemotherapy. The type and stage of the cancer, together with the patient's overall health, all have an impact on the treatment options. Ongoing research in skin cancer focuses on developing novel prevention methods and innovative treatment approaches. Vaccines that stimulate the immune system to target cancer cells are being investigated as a preventive measure. Similarly, targeted therapies, medications designed to selectively attack cancer cells while sparing healthy cells, are being explored. In addition to medical research, efforts to raise awareness about skin cancer and promote sun safety are of paramount importance. Educating the public regarding the risks associated with sun exposure and emphasizing the significance of protective clothing and sunscreen usage are crucial elements of public health campaigns. Successful initiatives have demonstrated a reduction in skin cancer incidence in certain populations. In conclusion, skin cancer is a pressing health concern necessitating ongoing research and prevention initiatives. By implementing effective strategies, it is possible to mitigate the incidence of skin cancer and enhance outcomes for those affected by this condition. Through collaborative efforts among healthcare professionals, researchers, and the public, substantial progress can be achieved in combating skin cancer.

Literature Survey

[1] Dorj, U. O et al. conducted a study on skin cancer detection using a combination of dermoscopy pictures and digital images. Their research involved the utilization of Artificial Neural Networks (ANN) as a classification algorithm, along with Support Vector Machine (SVM).

[2] A research journal from NIT Puducherry had performed a research on classification on skin cancer with model involving neural networks and SVM classifier. They tested the model with ISBI 2016. Among the algorithms they obtained accuracy of 86.23%. They embedded SVM classifier and obtained an accuracy of 88.02%.

[3] the geometric features from ABCD rule of melanoma, Grabcut image processing segmentation, and feature extraction applied to SVM classifier were used by Suleiman Mustafa, Ali Baba Dauda and Mohammed Dauda from National agency for science and engineering infrastructure obtained an accuracy of 80% on online skin diseases gallery with total of 200 images.

[4] on the ISIC 2019 dataset, selen Ayas proposed an approach for classifying skin disease using the swim transformer model and achieved an balanced accuracy of 82.3%.

[5] Using the ISIC dataset with the SVM and KNN classifiers, K.P Sanal kumar, A.Murugan and S.Anu H Nair published their work in the international journal of engineering and IJEAT and achieved a accuracy of 83.74% for SVM and 81.0 for KNN.

[6] In their study, the author introduced a novel approach called the Ballerina 5-layered K-Nearest Neighbors (KNN) Algorithm for classifying skin injuries. The proposed algorithm was tested on the PH2 dataset, a widely used dataset in dermatology research. The primary focus of the study was on the classification of melanoma, which is known to be more dangerous due to its relatively rapid spread and its potential to claim numerous lives. By leveraging the Ballerina 5-layered KNN Algorithm, the researchers aimed to improve the accuracy and efficiency of skin injury classification, particularly for the detection and differentiation of melanoma.

[7] In their study, Abbas Q et al. laid the groundwork for identifying skin diseases, specifically skin wounds with malignancy, through segmentation and the application of the K-nearest neighbors (KNN) algorithm on dermoscopy images. Melanoma, a dangerous form of skin disease responsible for 75% of melanoma-related deaths, was the focus of their research.

[8] Alum, MZ et al. proposed a method to identify melanocytes, a type of skin cell found in the epidermis layer, using KNN, Random Forest, and FUZZY C-means techniques. By overcoming this challenge, the authors aimed to enhance the diagnosis of melanoma, which is highly treatable if detected in the early stages but can be life-threatening if left undiagnosed.

[9] The author discussed the difficulties involved in analyzing skin lesion images, including issues such as light reflections and variations on the skin surface. They highlighted the lack of diversity in skin cancer classes and the insufficient availability of data in previous works. To address these challenges, Ramachandra Majji proposed a novel method that integrates gene expression data and machine learning algorithms for the detection of skin cancer.

[10] Binder examined a classification framework that was evaluated using unsegmented images. Cancer poses a significant threat to human life and can lead to unavoidable fatalities. With various types of cancer potentially present in the human body, accurate classification methods are crucial for timely diagnosis and treatment.

Data Set:

The dataset we used is International Skin Imaging Collaboration (ISIC) Skin cancer. It consists of 9 different classes of skin lesions . The images we used are high-resolution clinical photographs of skin lesions, taken with a variety of imaging modalities, including dermoscopy, clinical photography, and confocal microscopy. The size of skin cancer are:

-Dermatofibroma

- Actinic keratosis

-Basal cell carcinoma

the dermoscopic images are 400x600. We have resized them to 32x32 pixels inorder to get correct result and more accuracy. We have used 2,248 images for training. Dataset images are available in Kaggle website or at public ISIC-archive. The nine classes of

-Nevus

-Seborrheic keratosis

-Melanoma

-Pigmented benign keratosis

-Squamous cell carcinoma

-Vascular lesion

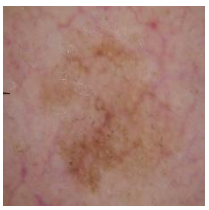
DATA AUGMENTATION

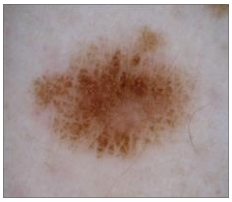
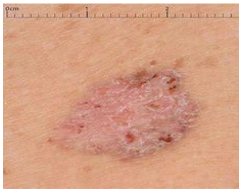
Our dataset is unbalanced we want to make it as balanced to increase the generalization of the model, and increase its accuracy. The goal of data augmentation is to generate new images that are identical to the original ones but different enough to provide additional training data for a machine learning model. After augmentation we have 1000 images in each class. The below table shows the data augmentation parameters that were used to form new images using the existing images

.

Data Augmentation Parameters	Value	Description
Rotation Range	20	Generates images with range -0.360
Width Shift Range	0.1	Rotates The image Horizontally in a Random Order
Height Shift Range	0.1	Rotates The image Vertically in a Random Order
Zoom range	0.1	Zoom-in or Zoom-out by 10%
Horizontal Flip	True	Fills the images horizontally for mirror reflection
Fill Mode	nearest	Fills the vacant place with nearest pixel value

Now the dataset is balanced. The datasets 9 classes are of same size. Thus, by testing the dataset it will not be underfitted.



Class	Original Training data	Augmented Training data	Original Testing data	Augmented Testing data
actinic keratosis	114	1000	16	200
basal cell carcinoma	376	1000	16	200
dermatofibroma	113	1000	16	200
nevus		1000	16	200
seborrheic keratosis	100	1000	3	200
melanoma	500	1000	16	200
pigmented benign keratosis	500	1000	16	200
squamous cell carcinoma	200	1000	16	200
vascular lesion	152	1000	3	200

Methodology

Image Pre-Processing:

Preprocessing is the stage when we get the data ready for Kera's model. Making the photos the same size and removing all null values from the dataset are the first action in data pre-processing. Additionally, we will split the data into training and testing sets. The dataset has undergone preprocessing, including image resizing, reshaping, and array form

cancer needs to be pre-processed. We have used open cv for resizing images. The Images are converted to 32x32 pixel size. Now those images are converted to grayscale image. Grayscale images are images that do not have any colour. In grayscale images, A single value that corresponds to the brightness or intensity of each pixel serves as its representation.

Data Augmentation:

The method of creating fresh image transformations using the provided image dataset increases the number of examples within the schooling set at the same time and introduces extra range in what the version sees and learns from.

Feature Extraction
GLCM(GRAY LEVEL CO-OCCURRENCE MATRIX) GLCM is

analysis, particularly in obtaining the distributed intensity from an object. It involves considering two pixel. GLCM yields several features, including contrast, correlation,

conversion. The images of skin Prominence, and shade. Each feature is calculated as follows:

Contrast: A specific formula is used to calculate the spatial frequency of texture in a skin lesion. Unfortunately, the formula for determining contrast in skin lesion texture analysis using GLCM has not been provided

Correlation: The following formula can be used to calculate a skin lesion's linear dependencies at different grey levels

$$\sum_{i,j}(i-j)^{2}c(i,j)$$

Energy: A skin lesion's degree of disorder or uniformity can be measured using the formula.

$$\frac{[\sum_{i,j}c(i,j)]-\mu_x\mu_y}{\sigma_x\sigma_y}$$

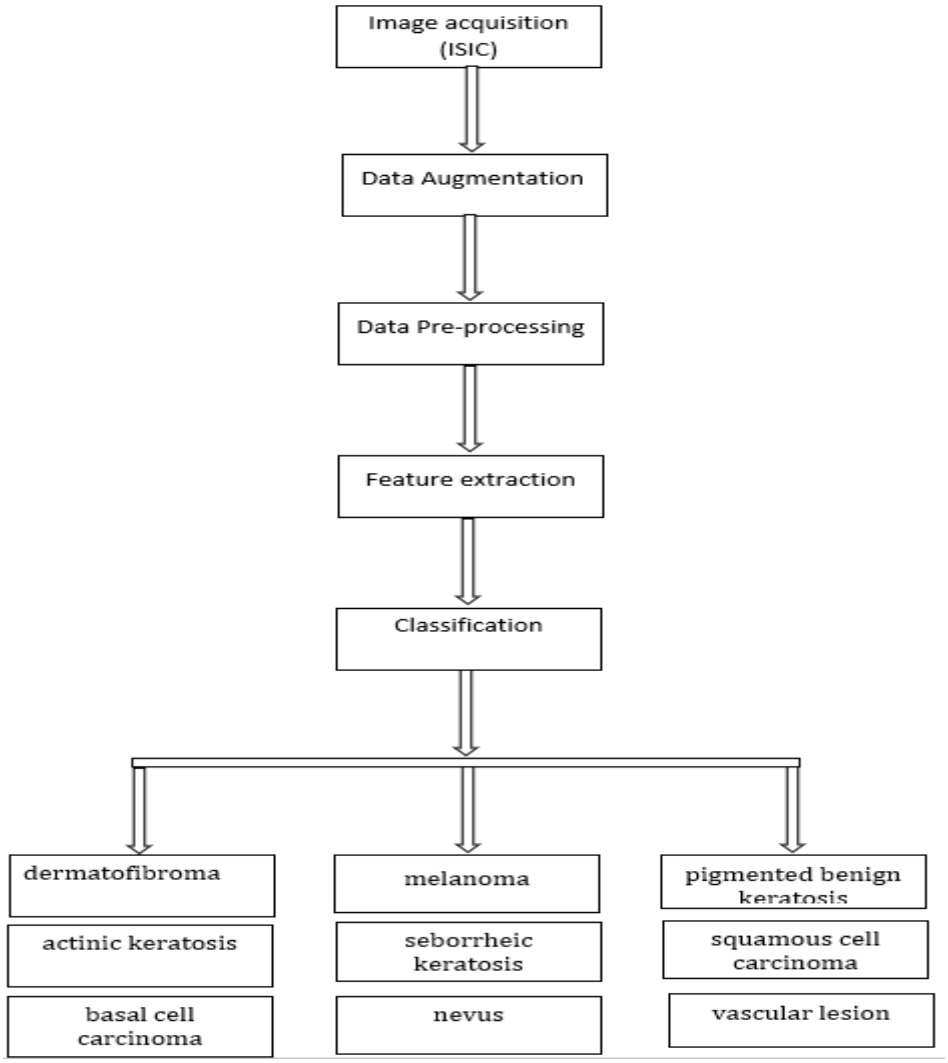
commonly employed for textural Homogeneity: The following formula can be used to determine how the elements are distributed throughout a skin lesion:

$$\sum_{i,j}c(i,j)\log(c(i,j))$$

energy, entropy, homogeneity,

Now we have to extract features from those images. Feature extraction involves the techniques like edge detection, histogram analysis, and texture analysis. Texture analysis refers to the process of capturing and quantifying various patterns, structures, and characteristics present in an image. Texture extraction involves analyzing the texture patterns within skin lesion images to extract relevant features that can aid in identifying potential cancerous regions. Skin cancer often exhibits distinct textural variations compared to healthy skin. By extracting texture features, such as variations in color, ASM, contrast, homogeneity and spatial arrangement of pixels, valuable insights into the characteristics of the skin lesion. Skin cancer lesions often exhibit specific color or intensity variations compared to healthy skin. Histogram-based techniques aim to quantify these variations and extract features for detection and diagnosis. The Gray Scale images are used for Histogram extraction. Edge detection is the method of detecting and highlighting the edges of an image. Edges can be defined as sharp discontinuities in image intensity or color. By highlighting these edges, we can extract important features from an image. Edge detection in skin cancer detection involves identifying and highlighting the boundaries or edges of skin lesions within an image. It plays a crucial role in image analysis and algorithms can identify irregularities or abnormal patterns that may indicate the presence of skin cancer. It is computed from GLCM matrix. Texture analysis involves the use of mathematical techniques to extract features from an image, such as the texture energy, coarseness, and contrast. Histogram analysis is the process of analyzing the distribution of pixel intensities in an image. A histogram is a representation of the pixel intensity distribution. In the context of skin cancer detection, histogram-based techniques are used to capture and analyze the intensity or color information present in skin lesion images. The pixel intensity or color values in an image can provide valuable insights into the characteristics of the skin lesion. Skin cancer lesions often exhibit specific color or intensity variations compared to healthy skin. Histogram-based techniques aim to quantify these variations and extract features for detection and diagnosis. The Gray Scale images are used for Histogram extraction. Edge detection is the method of detecting and highlighting the edges of an image. Edges can be defined as sharp discontinuities in image intensity or color. By highlighting these edges, we can extract important features from an image. Edge detection in skin cancer detection involves identifying and highlighting the boundaries or edges of skin lesions within an image. It plays a crucial role in image analysis and feature extraction for detecting potentially cancerous regions. Edge detection techniques aim to locate significant changes in pixel intensity or color values, which often correspond to the boundaries between different regions in an image. It is performed using Canny edge detection. Now we have to classify the images using various machine learning algorithms. Here we have used 4 types of algorithms. They are Decision Tree, KNN classifier, Random Forest algorithm, Voting Classifier. All these algorithms receive the same trained set of images. These models produce the output as which type of cancer it belongs to among those 9 different types

FLOW CHART



Random Forest

A popular ensemble learning approach for classification is random forest. The random forest method can be used to classify images into different groups based on their visual attributes in the context of skin cancer diagnosis using image data sets. A vast number of decision trees are created by the method, and each one is trained using a different random subset of the training data. Each decision tree is allowed to grow to a specific depth during training, which is set by a user-defined hyperparameter. The final prediction of the random forest algorithm is obtained by taking a majority vote over the predictions of all the decision trees. This approach has the significant benefit that it is applicable to both classification and regression issues. In the context of skin cancer detection, the random forest algorithm can be used to classify images into different categories based on their visual features. These features can be extracted using techniques such as edge detection, histogram analysis, texture analysis, and data augmentation. Once the features have been extracted, they can be used as input to the random forest algorithm. The features extracted from each image are used to determine the best split at each node of the decision tree. This process is repeated recursively until the tree reaches a user-defined maximum depth or until all the images at a given node belong to the same category. The final prediction is the category with the highest number of votes. Now random forest algorithm can be used to classify images into different categories such as melanoma, nevus, or seborrheic keratosis. By combining multiple decision trees, it is able to capture complex patterns in the data and achieve accuracy in classification tasks.

Voting Classifier:

Voting classifier is an algorithm of ensemble learning that brings together various machine learning algorithms to enhance classification performance .In the context of skin cancer detection, a voting classifier can be trained using various image data features such as edge detection, texture analysis, histograms, and data augmentation. By integrating the forecasts of various base classifiers trained on various subsets of the feature space, the voting classifier makes predictions. The voting classifier then takes a majority vote to make the final prediction. The goal of using a voting classifier is to reduce the bias and variance in the prediction by combining the predictions of multiple models. To build a voting classifier for skin cancer detection, we can first extract features from the image data using various techniques such as edge detection, texture analysis, and histograms. We also used data augmentation techniques to generate additional data for training and testing and improved the performance of the classifier. Next, we can train multiple base classifiers on different subsets of the feature space. For example, we can train a K-nearest neighbours classifier on the edge detection features, a random forest classifier on the texture features. Finally, we can combine the predictions of these base classifiers using the voting classifier algorithm. The voting classifier takes a majority vote of the predicted classes to make the final prediction. For example, if two base classifiers predict the sample as malignant and one predicts it as benign, the voting classifier will predict the sample as malignant. Overall, a voting classifier can be an effective approach for skin cancer detection as it combines the strengths of multiple machine learning models and improves the classification's overall accuracy.

KNN:

A machine learning method called K-Nearest Neighbour (KNN) can be utilised for categorization problems, including skin cancer detection using image data sets. The algorithm works by comparing a test image with the k-nearest training images and assigning the class label that is most frequent among the k-nearest neighbour. To use KNN for skin cancer detection using image data sets, we first need to prepare the data by extracting features from the images. This can be done using various techniques, such as histograms, edge detection and texture analysis. Once the features have been extracted, we can use them to train the KNN classifier. Now Train the KNN classifier on the training set by specifying Predict the class labels for the images in the test set by finding the k-nearest neighbour and assigning the class label that is most frequent among them. Analyse the KNN classifier's performance using metrics like accuracy, precision, recall, and F1 score. The performance of the KNN classifier can be improved by optimizing the value of k and by using feature selection techniques to select the most relevant features for classification. Overall, KNN is a simple yet effective algorithm for skin cancer detection using image data sets. However, it may not be as efficient as other algorithms for large data sets or high-dimensional feature spaces.

Decision Tree

A well-liked machine learning approach for detecting skin cancer is the decision tree. The algorithm functions by creating a model of decisions and potential outcomes that resembles a tree. A decision point is represented by each node in the decision tree, and potential outcomes are shown by the edges. The algorithm recursively divides the data into subsets based on the features, selecting the feature that best separates the classes at each node. In the context of skin cancer detection, the decision tree algorithm takes various features of skin lesions such as edge detection, histogram and texture as inputs and produces a decision tree based on these features. For example, the algorithm may divide the data into two subsets based on whether a lesion is symmetrical or not. If the lesion is symmetrical, it may further divide the data based on other features such as edge detection, histogram or texture. The decision tree algorithm is capable of handling both categorical and continuous data and can be easily interpreted by humans. It is also resistant to overfitting, which is a common problem in machine learning algorithms. However, it may not be suitable for complex data with many features, and small changes in the data can result in different decision trees. In summary, the decision tree algorithm is a useful tool for skin cancer detection, as it can effectively analyze multiple features of skin lesions and provide an interpretable output.

Evaluation and Experimental Results:

The effectiveness of our proposed model can be calculated using Accuracy, precision, recall and f1-score. Those metrics can be calculated with the help of TP, FN, FP, TN. The number of accurately anticipated positive images is known as True Positive (TP). The number of positively anticipated positive images is known as False Negative (FN). The number of incorrectly anticipated negative images is known as False Positive (FP). The number of accurately anticipated Negative images is known as True Negative (TN).

Confusion matrix:

An NxN matrix, where N is the projected number classes, is a confusion matrix.

Accuracy: the proportion of correct predictions made by a model among all predictions made. Precision: The Proportion of positive cases which are correctly identified

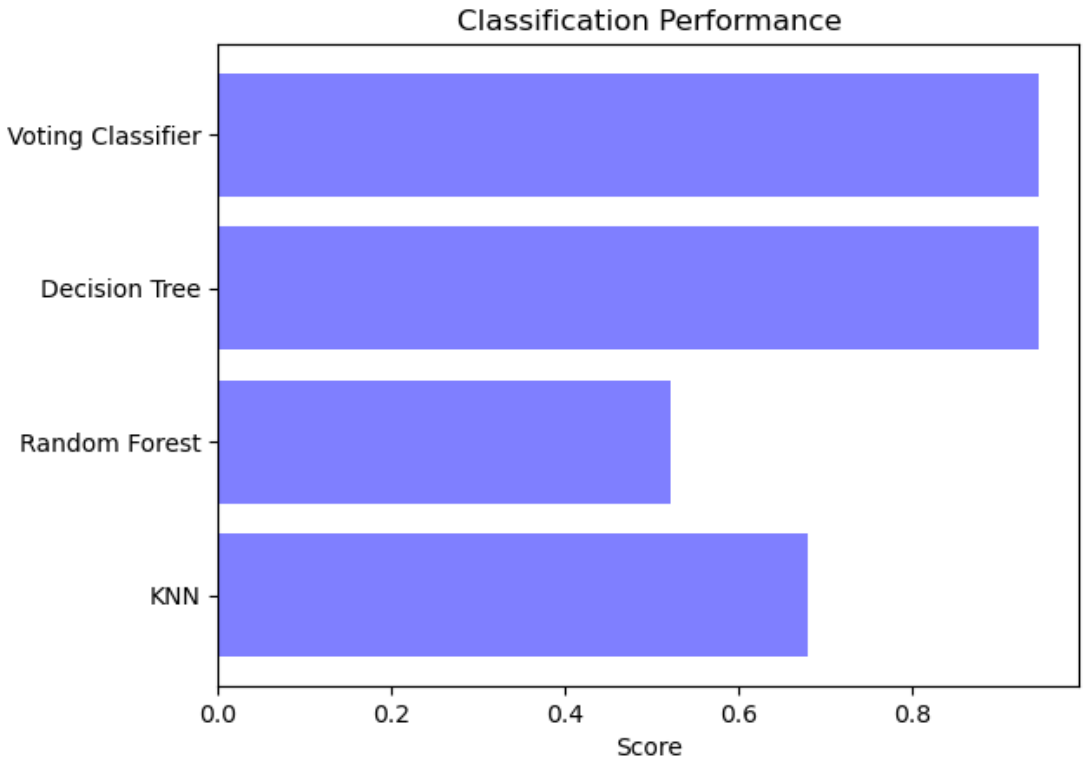
precision = TP / (TP + FP)

Recall: The proportion of Correctly Identified Actual Positive Cases recall = TP / (TP + FN)

F1-Score: For a classification task, the F1-Score is the harmonic mean of the precision and recall values.

F1-score = 2 * (precision * recall) / (precision + recall)

Model Name	Accuracy	Precision	Recall	F1- score
Decision Treee	93.8%	90	86	89
Voting classifier	93.8%	90	86	89
KNN	67.5%	60	60	58
Random Forest	50.6%	30	37	30



Conclusion

In this research, four models have been proposed and analysed. Out of them the Voting classifier and decision Tree performed well. The model was trained on (ISIC) International Skin Imaging Collaboration dataset, the best parameters setup were established after extensive research. According to the experimental results, It effectively distinguishes between different types of skin cancer. The effectiveness of the suggested model has been assessed and the model achieved a 93.8% accuracy rate

References:

[1] The automatic identification of melanoma by wavelet and curvelet analysis: Study based on neural network classification, M. K. A. Mahmoud, A. Al-Jumaily, and M. Takruri, 2011 11th International Conference on Hybrid Intelligent Systems (HIS), pp. 680–685.

[2]Niharika Hegde, M. Shishir, S. Shashank, P. Dayananda, and Mrityunjaya V. Latte, "A Survey on Machine Learning and Deep Learning based Computer Aided Methods for Detection of Polyps in CT Colonography", Current Medical Imaging, vol. 17, no. 1, pp. 3–15, 2021.

[3] "Multi-level Attentive Skin Lesion Learning for Melanoma Classification", IEEE Engineering in Medicine & Biology Society (EMBC), 43rd Annual International Conference, 2021, pp. 3924-3927. By X. Wang, W. Huang, Z. Lu, and S. Huang

[4]"Structural Basis of Skin Melanoma Metastasis", 2022 Ural-Siberian Conference on Computational Technologies in Cognitive Science Genomics and Biomedicine (CSGB), pp. 340-342, by N. Bgatova, A. Lomakin, I. Taskaeva, N. Obanina, V. Makarova, and A. Letyagin.

[5]"Melanoma Detection Using Deep Learning", 2022 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-4, by K. Padmavathi, H. Neelam, M. P. K. Reddy, P. Yadlapalli, K. S. Veerella, and K. Pampari.

[6]"Lesion Attributes Segmentation for Melanoma Detection with Multi-Task U-Net", 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 485-488. E. Z. Chen, X. Dong, X. Li, H. Jiang, R. Rong, and J. Wu.

[7] Materials Today: Proceedings, vol. 24, pp. 241-250, 2020. E. C. Yuvaraju, L. R. Rudresh, and M. Saimurugan. "Vibration signals based fault severity estimation of a shaft using machine learning techniques.".

[8] S. Jp, M. Kumar, and K. P. Soman, "Deep learning-based techniques to enhance the precision of phrase-based statistical machine translation system for Indian languages", International Journal of Computer Aided Engineering and Technology, vol. 13, p. 239, 2020.

[9]Electrical Engineering (ICEE) Iranian Conference on, S. A. Mahdiraji, Y. Baleghi, and S. M. Sakhaei, "BIBS a New Descriptor for Melanoma/Non-Melanoma Discrimination", pp. 1397–1402, 2018.

[10] J.-A. Almaraz-Damian, V. Ponomaryov, S. Sadovnychiy, and H. Castillejos-Fernandez, "Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures", Entropy, vol. 22, no. 4, pp. 484, Apr. 2020.

[11] Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, pp. 4943-4955, 2022, ISSN 1319-1578. Suraj Kotecha, Adithya Vasudevan, V.M.K. Kashyap Holla, Satyam Kumar, Dayananda Pruthviraja, and Mrityunjaya Vithal Latte.

[12] "Skin disease classification versus skin lesion characterization: Achieving robust diagnosis using multi-label deep neural networks", 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 355–360. Haofu Liao, Yuncheng Li, and Jiebo Luo.

[13] M. A. Khan, M. Y. Javed, M. Sharif, T. Saba, and A. Rehman, "Multimodel deep neural network based features extraction and optimal selection approach for skin lesion classification", in Proc. Int. Conf. Comput. Inf. Sci. (ICCIS), pp. 1-7, Apr. 2019.

[14] "Simple and effective pre-processing for automated melanoma discrimination based on cytological findings", 2016 IEEE International Conference on Big Data (Big Data), pp. 3439-3442. T. Yoshida, M. E. Celebi, G. Schaefer, and H. Iyatomi.

[15] Skin Cancer Detection Using Convolutional Neural Network, 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0169-0176. D. C. Malo, M. M. Rahman, J. Mahbub, and M. M. Khan.

[16]"Classification of ECG Signal Using Machine Learning Techniques", Proceedings of the 2019 2nd International Conference on Power and Embedded Drive Control ICPEDC

2019, pp. 122-128, S. S., G. S. Sweta, P.I. K., K. and K. J. Mohan Reddy, 2019.

[17] "A Fuzzy-Neuro-Based Clinical Decision Support System For Disease Diagnosis Using Symptom Severity" by S. Tandra, D. Gupta, J. Amudha, and K. Sharma is published in Advances in Intelligent Systems and Computing, Springer, vol. 1118, 2020, pp. 81–98.