Skin Cancer Recognition and Detection Using Machine Learning Algorithm

A. Jenitha, G. Amrutha, K.L. Kishore, K.R. Rohan, S.N. Sagar

Dr. T. Thimmaiah Institute of Technology

Visveswaraya Technological University

KGF, Karnataka India

jenitha@drttit.edu.in

Abstract. In this paper, we concentrate on the identification of skin cancer. The skin images are taken from a medical database which is a pre-processed image, which is given as input for different machine learning algorithm. The algorithm used is KNN classifier, SVM classifier, and CNN model. where these classifiers will classify whether a given image is cancerous or non-cancerous image. In case of the KNN and SVM the output is 80%, hence in CNN model substantial improvement in accuracy of cancer detection is obtained & it can classify the cancerous & Non-cancerous images efficiently. The process was conducted for test data, training data and validation data using different-images. The training dataset was trained with 100 epochs. The process obtained the accuracy of 97% in training result. in testing result obtained is 95% of accuracy and 96% for validation testing.

***Keywords*:** Skin Cancer, CNN, Melanoma, SVM, KNN, Machine Learning.

1 Introduction

In this project we try to use appropriate machine learning methods to form an efficient system for melanoma detection. For this process, the input-images are taken from medical institute. It consists of collection images taken from medical institute consisting of dataset about 10000 images of 3Gb [1]. Which contains variety of cancerous and non-cancerous images collected from different patients and it's a processed image examined by doctors across the world [2]. We have taken the images from medical website Kaggle [3]. The results show that the proposed system can out-perform previous methods. The remaining part of this paper is given as follows. In section II, the related works of this project is explained in detail. In section III, the proposed scheme is completely explained. Results of our method are given in section IV. section V concludes the paper.

ICASISET 2020, May 16-17, Chennai, India

DOI 10.4108/eai.16-5-2020.2304046

2 Related Works

RinaRefianti et al [1] proposed “Classification of Melanoma Skin Cancer using Convolutional Neural Network” In this paper they used deep learning technology with CNN method. The result of training with different number of training and epochs resulted the percentage of 93% in training and 92% in testing. JaineshRathod et al, [2] proposed “Diagnosis of skin diseases using Convolutional Neural Networks” In this paper, they proposed automatic image-based system for recognition of skin diseases using ML classification. The classification gives the output accuracy of 70%. ShekharYadav, et al from MM University of Technology Gorakhpur, India [3] proposed “Melanoma Skin Cancer Detection Using Various Classifiers” In this paper, they proposed the system of gathering dermoscopy picture database. Classification consists of SVM, KNN, decision trees and ensemble classifiers. The highest accuracy obtained from all classifiers was for SVM model and it was more than 90%. DiwakarGautam, et al department of CSE and MNITE [4]. Vinod Jagannath Kadam et.al,[6][7] proposed “Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Auto encoders and Soft max Regression”.

3 Proposed Method

Skin cancer detection using machine learning algorithm can be implemented by using KNN/SVM/CNN algorithm. KNN algorithm is used and optimized for the process of design and research [5]. After performing the KNN algorithm we got 79% accuracy. After performing with the SVM algorithm we got 80% accuracy. Convolutional neural network is a deep learning algorithm which is capable of taking an input image and assigning importance (learnable weights and biases) to special objects in an image and to segregate it from other. The flowchart of the project is given in figure 1a.

Database- It consists of collection images taken from medical institute consisting of dataset about 10000 images of 3Gb. Which contains variety of cancerous and non-cancerous images collected from different patients and it’s a processed image examined by doctors across the world. From the website called kaggel. The sample train x and train y images given in figure 1b.

Division of data into Train Data and Test Data- Dividing the data-set into training set and test set to examine our model performance the data is spitted into train data and test data, where the train data is analyzed through different procedures and then finally estimated with the test data. when we plot an image to see how the original images look, we get plot of original images and masked images shown in figures 2a and 2b respectively.

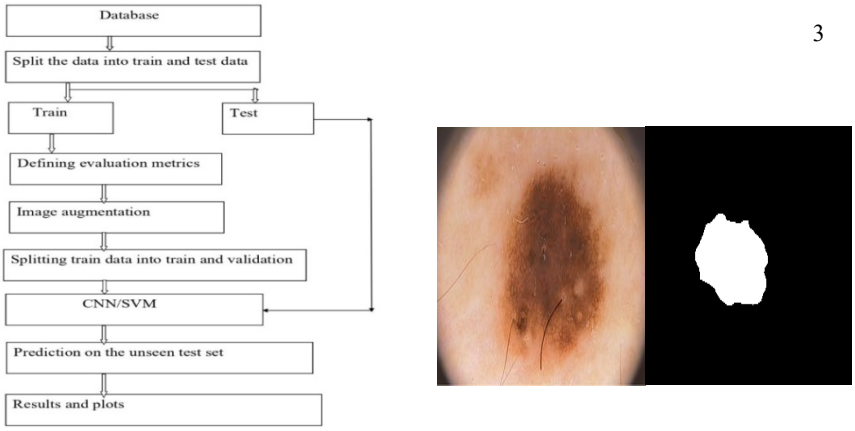

Fig 1a.Flowchart of the Project and **Fig 1b.**Train x and Train y Images.

Evaluation Metrics

Intersection over Union- Measures similarity between finite sample sets, and is given as the size of the intersection by the size of the union of the sample sets: it's defined as IOU,

$$(\text{IOU}) = \frac{\text{AREA OF OVERLAP}}{\text{AREA OF UNION}} \quad (1)$$

Dice Co-efficient- represent in Eq. 2

$$\text{DICE} = \frac{2\times TP}{(TP+FP)+(TP+FN)} \quad (2)$$

Precision- To determine, whether the costs of False Positive is high. It's given as

$$\text{PRECISION} = \frac{TP}{TP+FP} \quad (3)$$

Recall- Determines the actual positives the model can capture through labeling it as positive (True Positive). $\text{RECALL} = \frac{TP}{TP+FN}$ (4)

Accuracy- Accuracy is measured in terms of the given by the formula

$$\text{ACCURACY} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (5)$$

Image Augmentation - Artificially creating training images through various ways of processing/combining multiple processing techniques like random rotation, shifts, shear and flips. It's shown in figure 3a.

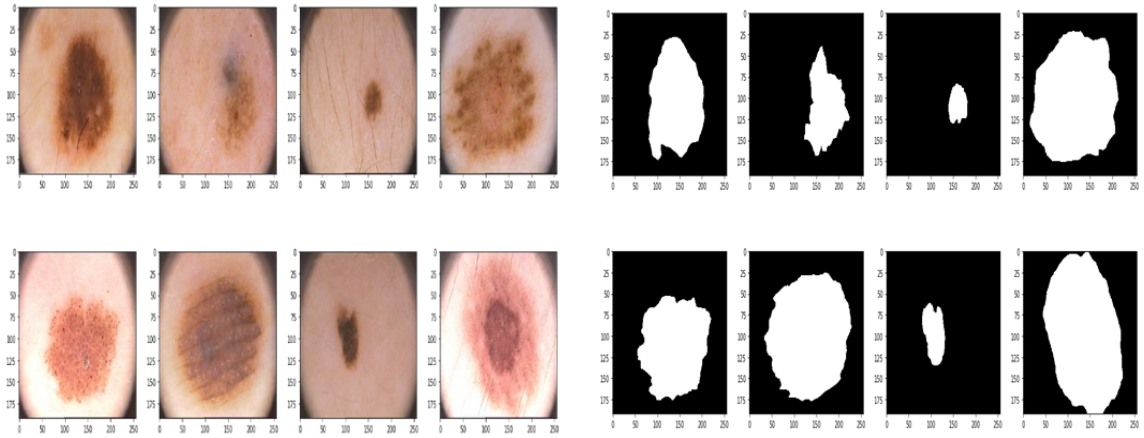

Fig 2a. Plot of Original Images and **Fig 2b.**Plot of Masked Images.

Making a Validation Set- Now we join all the augmentations image arrays to the original training arrays. We will split our full training set into train and validation set. Validation dataset is used to validate the performance after each epoch.

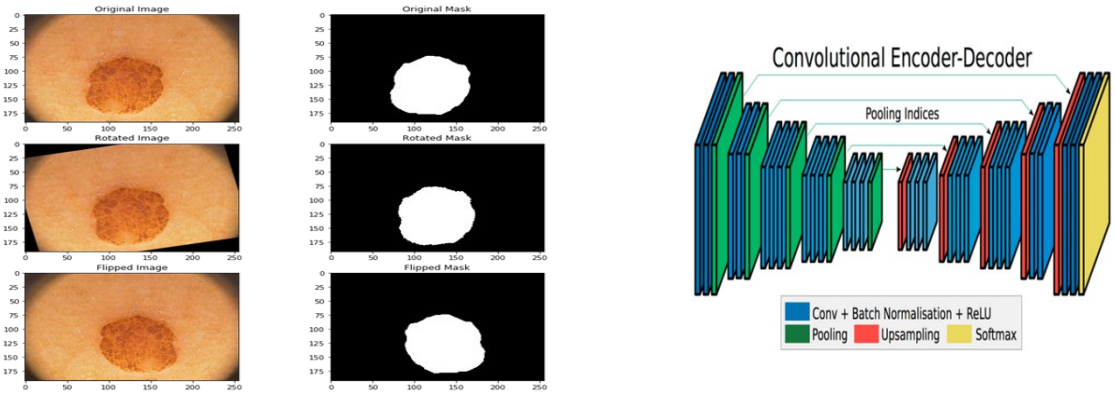

Fig 3a. Image Augmentation and **Fig 3b.**Convolutional Encoder-Decoder

Implementation Model- The proposed work using CNN is discussed below, normally we should have maximum number of samples for training of any CNN method. We're using a dataset containing of 10,000 melanoma and dermoscopic images for detection of melanoma from non-dermoscopic images. The network is trained through 100 epochs to get maximum accuracy. The following is the values used for epochs. Length of the Training Set: 360, Length of the Test Set: 50 and Length of the Validation Set: 90. Architecture of the proposed model is illustrated in figure 3b

Formulae Used in SGD algorithm

Convolution operation can be written with the formula given as follows:

$\mathbf{s(t) = (x \times w)}(6)$Where, s(t)= Feature Map, x= input and ω= kernel. [1] Convolution operations with more than 1-D input can be given as follows

$\mathbf{s(i,j) = (K \times I)(i,j) = \sum m \sum n\, I(i-m, j-n) K(m,n}$ (7) Where, i and j= pixels of the image, k= kernel, I= input. [2]

Initialization method can be given as$\mathbf{W_{i,j=U\left[\frac{-1}{\sqrt{-n}},\frac{1}{\sqrt{n}}\right]}}$ (8).

Where, W= weights at each layer, U= uniform distribution of a particular interval and n= Layer's size. [3]

Goal of the learning method is to get a value of weights that minimizes the LF: $\mathbf{W_{k=}W_{k-1} - \epsilon \frac{dE(w)}{dw}}$ (9). where, WK= weight to be updated, WK−1= previous weight, epsilon= learning rate and dE(W)/dW= derivative of loss or error function. [4] Hyper-parameter: The expression is given below where, W is the volume size, F is filter, S as stride applied, P as number of zero padding used [5]

$\mathbf{H = \frac{W-F+2P}{S+1}}$ (10) Relu- The expression is given below, where the x is the input image given to ReLu input layer, it is denoted as f(x) [6] $\boldsymbol{f(x) = ln(1 + e^x)}$ (11)

4 Experimental Results

Section of training data, test data and validation set, where length of training data was 360, length of test data was 50 and length of validation set is 90. Each section of training data was trained for hundred epochs. We use googleCollaboratory to run our project code which is written in python programming language. The readings after one complete epoch is given as in the form of loss, IOU, precision, recall and accuracy. And the complete values of the training data, test data and validation data are also given, we've obtained an accuracy of 97.35% for the training set, 94.66% for the test set and 96.01% for the validation set, we've obtained higher accuracy for all the sets of data and the normal accuracy of 97% obtained after 100 epochs meets our project objective of obtaining higher accuracy.

Plots for Training Statistics on the Train Set- The plot clearly shows that the loss at the initial stage was high and accuracy was low and after the completion of 100 epochs, the loss has gradually decreased to a very low value and accuracy has increased to a higher value. It's shown in figure 4.

Prediction on the Test Data- The prediction on the test data is carried out by plotting on the test data of length 50, we take 5 random images and plot its original image and compare ground truth of it and our predicted output. Predicted output is almost (97%) closer to the ground truth image, with only minor losses obtained. So, we can clearly say that the prediction on the test data is successful and the output is almost accurate to the original image.

Comparison of the Prediction after Enhancement- We now compare our prediction after enhancing the image, we can see that our output obtained is accurate and close to 97% accuracy and the objective of obtaining higher accuracy for detecting skin cancer is obtained.

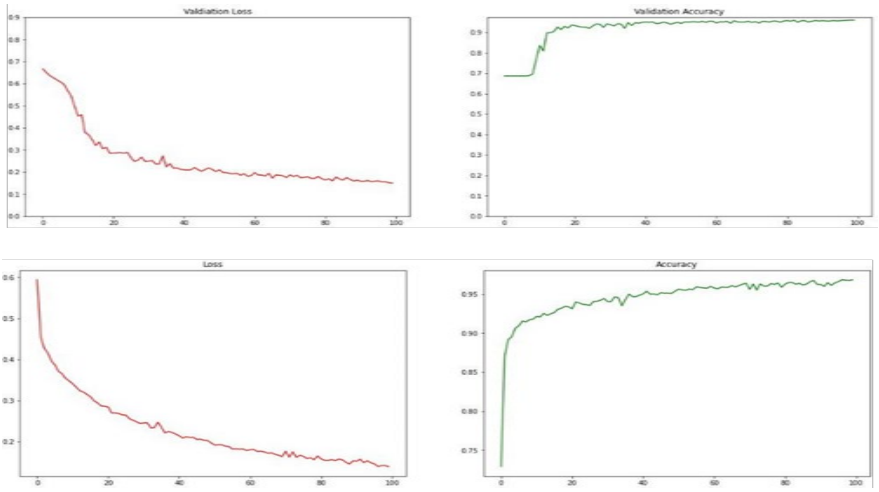

Fig 4 Plot of Training Characteristics

5 Conclusion and Future Work

The process was implemented on images of test data, training data and validation data using different-images and for hundred epochs in the period of training. The process obtained the maximum accuracy of 97% for training result, testing result was 95% of accuracy and 96% in validation testing. Where the section of training data, test data and validation set, where length of training data was 360, length of test data was 50 and length of validation set is 90. All training data was trained with hundred epochs.Obtained results shows that the length of training data and epochs for training determines the amount of accuracy in classifying. The training of more data, the better the results can be produced. Hundred epochs are the ideal epoch to obtain the best accuracy. For future work, if clustering method is implemented for our project the present stage of the skin cancer can be detected. Future works should try to use stand-

ard available protocols and established methods used for process of the project, to have compatibility and to reduce loss as much as possible.

References

1. Skincancer.org, “Melanoma - SkinCancer.org,” 2016. [Online].
2. American Cancer Society, “Cancer Facts & Figures 2016,” American Cancer Society, Atlanta, GA, USA, 2016.
3. Ahmed, S.T., Sandhya, M. & Sankar, S. TelMED: Dynamic User Clustering Resource Allocation Technique for MooM Datasets Under Optimizing Telemedicine Network. Wireless PersCommun 112, 1061–1077 (2020). https://doi.org/10.1007/s11277-020-07091-x
4. S. T. Ahmed and S. Sankar, "Investigative Protocol Design of Layer Optimized Image Compression in Telemedicine Environment", Procedia Computer Science, vol. 167, pp. 2617-2622, 2020, [online] Available: https://doi.org/10.1016/j.procs.2020.03.323
5. Gunashree, M., Ahmed, S. T., Sindhuja, M., Bhumika, P., Anusha, B., &Ishwarya, B. (2020). A New Approach of Multilevel Unsupervised Clustering for Detecting Replication Level in Large Image Set. *Procedia Computer Science*, *171*, 1624-1633.https://doi.org/10.1016/j.procs.2020.04.174
6. Vinod Jagannath Kadam, Shivaji rao Manik rao Jadhav, K.Vijayakumar, “Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Auto encoders and Soft max Regression”, Image & Signal Processing, springer, june 2019.
7. K. Vijayakumar , K. Pradeep Mohan Kumar ,Daniel Jesline, “Implementation of Software Agents and Advanced AoA for Disease Data Analysis”, journal of medical systems, Part of Springer Nature 2019.