IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SKIN CANCER DETECTION USING MACHINE LEARNING TECHNIQUE

Mansi Mishra
Mtech Scholar
Dept of CSE , SSTC Bhilai

Dr. R.K. Khare
Assistant Professor
Dept of CSE , SSTC Bhilai

Abstract:

Skin disorders are a prevalent public health issue in every region of the world. It is impossible to see the dangers posed by the diseases, which not only wreak havoc on one's physical health but also set the stage for mental sadness. In addition to this, it has been linked to incidences of skin cancer in extreme circumstances. As a consequence of this, one of the most difficult jobs involved in medical image analysis is the diagnosis of skin disorders based on clinical photographs. In addition, identifying skin illnesses manually by medical professionals is an arduous and time-consuming process that is also very subjective. As a direct consequence of this, patients and dermatologists alike want automated skin disease prognosis, which allows for more efficient treatment planning. In this study, we provide a method for the removal of unwanted hair from digital photographs that is based on morphological filtering, namely the Black-Hat transformation and the inpainting algorithm. Following this, we use Gaussian filtering to deblur or denoise the images. In addition, we use an automated ANN segmentation approach to separate the healthy lesions from the ones that are impacted. We use a technique called the Grey Level Co-occurrence Matrix (GLCM) in conjunction with statistical characteristics in order to extract the underlying input patterns from the skin photos. Three computationally efficient machine learning techniques, Decision Tree (DT), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) classifiers are applied using the extracted features for effectively classifying the skin images as melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), vascular lesion (VASC), and Squamous cell carcinoma (SCC). the HAM10000 datasets are utilized in the process of validating the models. SVM has a performance that is marginally superior to that of the other two classifiers. In addition to this, we have evaluated our procedures in light of the most recent scientific developments.

Keywords: Skin Disease, Machine Learning, Image Segmentation, Decision Tree, Support Vector Machine, Nearest Neighbor (KNN)

1. Introduction

There are a number of organs that make up the human body. The human epidermis is one example. It is the biggest organ in the human body and covers the whole of the human body [1]. The term "skin disease" refers to any condition that manifests on human skin [2]. Skin disorders are among the most infectious illnesses that may be spread from person to person. About 2,794 people in Bangladesh lost their lives to skin cancer in 2018, according to data provided by the World Health Organization (WHO) [3]. According to the World Health Organization (WHO), more than 14 million people were diagnosed with the disease in 2018, and 9.6 million people lost their lives as a result of it [4]. It is characterized by a shift in the skin's color or texture. Skin illnesses can have a variety of origins, including bacterial infections, viral infections, allergic reactions, and fungal infections [5]. In addition, skin diseases can be caused by the hereditary factor. In most cases, diseases of the skin manifest themselves in the epidermis, which is the thin outermost layer of the skin and is visible to the human eye. These conditions can cause mental anguish and ultimately result in bodily harm. Actinic keratosis (AK), Basal cell carcinoma (BCC), Benign keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic nevus (NV), Squamous cell carcinoma (SCC), and Vascular lesion (VASC) are the different forms of skin lesions that are depicted in Fig. 1. The lesions present a variety of symptoms and range in terms of their degrees of severity. Some of them are permanent, while others are only transitory and may or may not cause any discomfort. Melanoma is the most lethal and devastating form

of skin disease among all of these skin conditions. However, if treatment is started early enough, around 95 percent of people with skin diseases can make a full recovery. When it comes to precisely classifying skin illnesses, an automatic computer-aided approach can be of great assistance. There is a significant knowledge gap between dermatologists and patients who suffer from skin diseases since many individuals are unaware of the many kinds of skin diseases, their symptoms, and the phases at which they progress. Sometimes it takes a considerable amount of time for the symptoms to become apparent. In order to do this, early and prompt identification is required. On the other hand, accurate diagnosis of skin illnesses can be challenging and costly, as it requires determining both the kind of disease and its stage. The machine learning-based autonomous computer-aided system can identify skin problems faster and more accurately.. Over the past three decades, a significant amount of work has been done by researchers to classify skin diseases. The region is so significant that it has emerged as a prevalent focus of study interest. There has been a lot of study done on the identification and categorization of skin diseases, but there is still a hole in the field that needs to be addressed. The majority of the work that has been done in the past is based on a single disease [6, 7], and the work that has been done is insufficient for categorizing more than one category [8]. The work of classifying many classes is particularly difficult since the symptoms of skin diseases tend to be very similar to one another.

The following is an overview of the primary contribution that this research has made:

- Construct a model for dehairing by utilizing the Black-Hat Transformation and Image In painting method.
- To design a strong segmentation model utilizing the Grab cut approach that can detect the lesion without compromising any of the picture's information and make the image more suited for additional processing.

• To build an automated classification model for skin disorders that has a high level of accuracy and is based on a sufficient number of relevant characteristics.

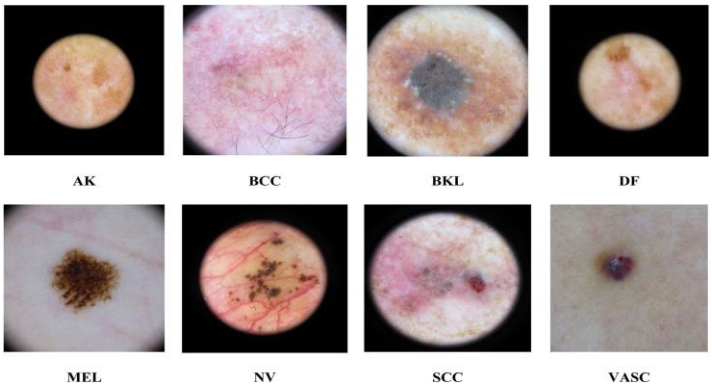


Fig. 1. Dermatology Sample Ham 10000 Challenge Dataset images.

The remaining parts of the investigation are laid up as follows: In the second half, we will discuss the relevant literature as well as the problem statement. Part 3 provides a description of the approach that has been proposed, which includes preprocessing, skin lesion segmentation, feature extraction, and skin disease classification. Part 4 covers the topics of the datasets that were utilized for this study as well as the data preparation procedures. Part 5 contains a discussion on the experimental findings that support the suggested methodology. This chapter includes contains a quantitative analysis as well as a discussion on how to evaluate performance. Part 6 brings the whole thing to a close by discussing potential outcomes.

2. Literature Review

Many researchers suggested skin disease categorization methods. Datasets, feature extraction, feature selection, and classification models classify related works. This section reviewed relevant research publications to find prior studies' methods and gaps.

Jagdish et al. [9] suggested an image processing-based skin disease diagnosis model. They used KNN and SVM classification algorithms with wavelet analysis to fuzzy cluster 50 sample photos. The K-Nearest Neighbor classification algorithm outperformed the Support vector machine (SVM) classification methodology with 91.2% accuracy and correctly recognized the skin illness. They used just 50 photos of basal and squamous disease.

Naeem et al. [10] used image processing and SVMs to predict skin cancer. They employed GLCM to isolate image highlights after pre-handling for noise removal and picture enhancement. Finally, the classifier rated the photos' safety.

Bandyopadhyay et al. [11] suggested a DL-ML model. They used Alexnet, Googlenet, Resnet50, and VGG16 for feature selection and Support Vector Machine, Decision tree, and Ensemble boosting Adaboost classifier for classification. Finally, they conducted a comparison research to determine the best prediction model. Kalaivani et al. [12] introduced a new method that integrates two data mining procedures into a single unit and an ensemble approach that groups both methods together. They categorized skin problems into seven groups using ensemble deep learning on an informative Dermatology dataset ISIC2019 picture. The ensemble method predicted skin disorders better.

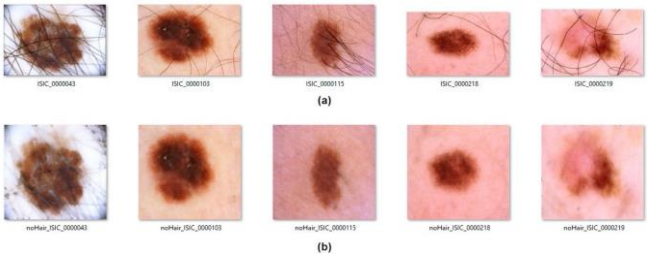
AlDera et al. [13] created a skin imaging model to identify acne, cherry angioma, melanoma, and psoriasis. Dermnet NZ and atlas dermatologico employed Otsu's image segmentation, Gabor, Entropy, and Sobel feature extraction. Finally, SVM, RF, and K-NN classifiers yielded 90.7%, 84.2%, and 67.1% accuracy.

Kshirsagar et al. [14] proposed MobileNetV2/LSTM skin disease categorization. They prioritised skin disease prediction and state data efficiency. The texture-based model beat CNN and FTNN in skin sickness detection and cancer development assessment.

Hatem [15] shown how to identify benign skin lesions. He identified skin lesions using k nearest neighbour (KNN) classifier and achieved 98% accuracy with just two classifications.

CNN-based skin disease classification was proposed by Kethana et al. [16]. 10015 ISIC 2019 photos identified Melanoma, Nevus, and Sebborheic Keratosis with 92% accuracy.

Yao et al. [17] presented a single-model skin lesion classification technique for limited and imbalanced datasets. They trained Deep Convolutional Neural Networks (DCNN) on short, unbalanced datasets, applied regularisation DropOut and DropBlock to avoid overfitting, and suggested a Modified RandAugment augmentation approach to address sample underrepresentation in the small dataset. Finally, a novel Multi-Weighted New Loss (MWNL) function and an end-to-end cumulative learning strategy (CLS) reduced aberrant sample impacts on training and addressed unequal sample size



and classification difficulty. This model classifies well in a short computational and inference time.

Padmavathi et al. [18] suggested a pre-trained and fine-tuned deep learning network for skin lesion classification. They compared performance utilizing transfer learning approaches and quantitative measures including specificity, sensitivity, precision, and accuracy.

Maduranga et al. [19] developed an AI-based skin disease screening smartphone app. They used CNNs to classify skin diseases using HAM10000 dataset. MobileNet with transfer learning was used to build a fast and accurate identification mobile app with 85% accuracy.

Jain et al. [20] developed an optimum probability-based deep neural network (OP-DNN) to identify skin diseases. After removing undesired information from photos, they extracted characteristics to train Optimal Probability-Based Deep Neural Networks (OP-DNN). Optimization yielded 95% accuracy, 0.97 specificity, and 0.91 sensitivity.

3. Proposed Methodology

This chapter explored skin disease classification methods. The procedure includes: Image preprocessing, segmentation, feature extraction, and classification follow. Fig. 2 shows a strategy overview.

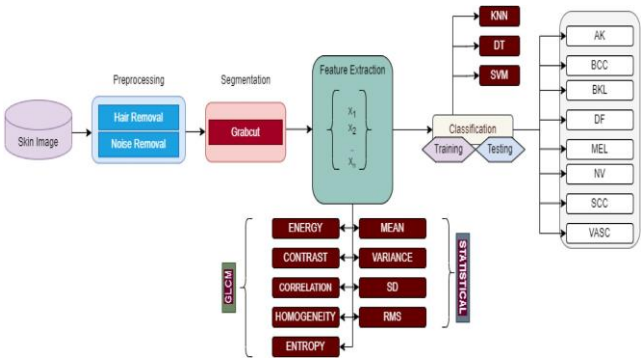


Fig. 2. A Generalized Block Diagram of the Skin Diseases Classification System

3.1. *Image Preprocessing*

Preprocessing an image improves it. Skin photos may include undesired hair, noise, or distortion. Image quality affects image processing performance. Processing skin photos improves skin disease detection system performance. It enhances picture quality, simplifies processing, and boosts accuracy. Image Resizing, Hair Removal, and Noise Removal include preprocessing.

3.1.1. *Image Resizing*

Features vary by picture size. To fix this, input photos are resized. It speeds processing and boosts system performance. We scaled all supplied photos to 512 × 512. Figs. 3(a) and 3(b) show the original and scaled images.

Fig. 3. The figure shows image preprocessing (a) Before Resizing (b) After Resizing.

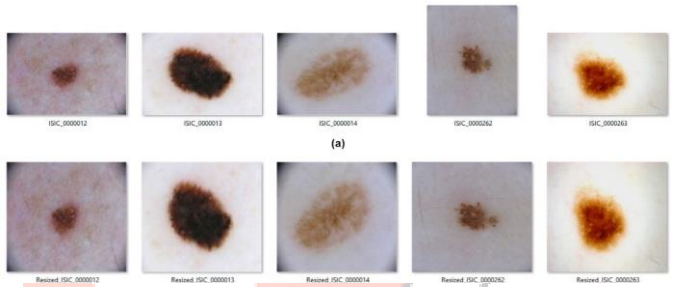


Fig. 4. Preprocessing of Images (a) Before Removing Hair (b) After Removing Hair.

3.1.2. Hair Removal

Hair removal improves skin disease detection. Hair removal from skin photos was done using median filter [34], adaptive threshold, Gabor filtering, PDE-based, and morphological processes like Top Hat filter. Our suggested technique removes hairs using morphological filtering like Black-Hat transformation and inpainting algorithms. DHR algorithm steps: 1. Grayscale RGB pictures. 2. Morphological Black-Hat grayscale pictures. 3. Make an inpainting mask. 4. Apply inpainting algorithm to original picture using this mask. Figs. 4(a) and 4(b) show before and after hair removal.

3.1.3. Noise Removal

Digital picture collection, transmission, and processing can cause noise. Unexpected brightness or color changes decrease image quality. Mean, median, Gaussian, and bilateral filtering were used to blur pictures or remove noise. Gaussian filtering blurs and removes noise in this system. It convolutional processes pictures. (1) The kernel values have a Gaussian distribution..

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{(Eq.1)}$$

3.2. Segmentation

Picture segmentation refers to the act of separating a picture into distinct groups or sections such that they do not overlap one another. It takes into account aspects such as the amount of grey, the brightness, the color, the contrast, and the texture. It differentiates the same lesions from the healthy skin that is located surrounding them. When it comes to accurately evaluating photos, this stage is the most important one since it determines how accurate the procedures that come after it will be. However, due to the significant variances in size, shape, and color of the lesions, correct segmentation in microscopic pictures can be a difficult task. In addition to this, there is a lack of contrast between the healthy skin around the lesions and the lesions themselves. Image segmentation may be broken down into many categories, including threshold-based, region-based, cluster-based, and edge-based [40–43]. Each of these categories has a variety of different segmentation approaches. In this study, we used the automated ANN approach for picture segmentation using k-means clustering and the Hue Saturation Value (HSV) color space. This was done in order to get the best possible results. We chose to utilize the ANN approach because it provides an estimate of the color distribution of both the target item and the backdrop by employing a Gaussian Mixture Model (GMM). This model automatically generates a rectangle and separates the foreground from the background photos, hence reducing the amount of human engagement that is required.

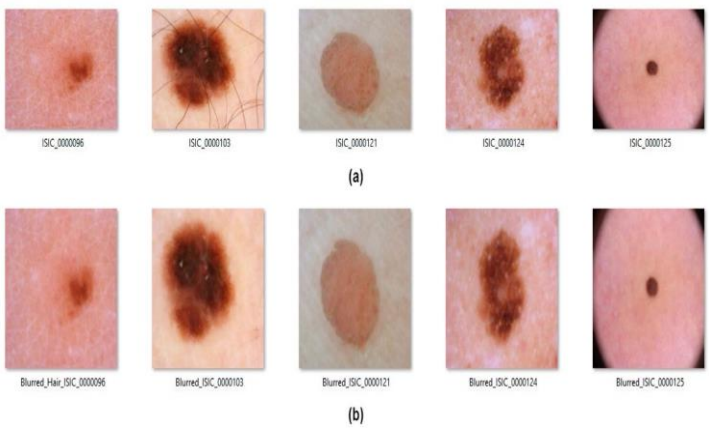


Fig. 5. Preprocessing images before and after Gaussian filtering.

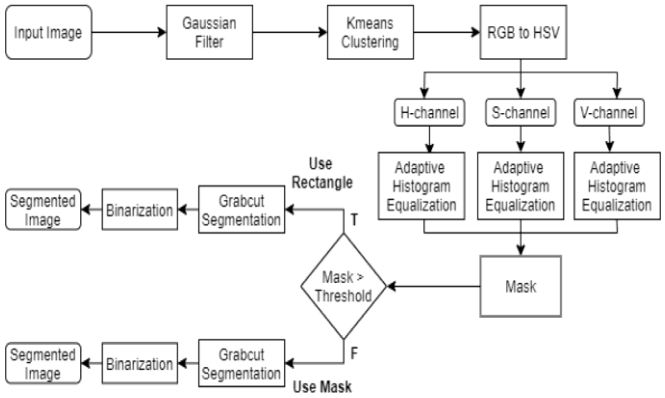


Fig. 6. Block diagram of automatic ANN segmentation.

4. Dataset

1.1. Dataset Description

We trained, tested, and assessed the proposed model using two well-known datasets: the "International Skin Imaging Collaboration (ISIC) 2019" challenge dataset and the "HAM10000 (Human-Against-Machine with 3000 training photographs)" dataset. Both datasets compared human and machine performance.

4.1.1. Dataset Preparation

The HAM10000 dataset is similarly an uneven dataset, with the AK class including 327 photos, the BCC class containing 514 images, the BKL class containing 1099 photographs, the DF class containing 115 images, the MEL class containing 1113 images, the NV class containing 3000 images, and the VASC class containing 142 images. Figure 8 presents the data distribution that pertains to each class for the ISIC 2019 dataset. During the process of training the model with the unbalanced dataset, there is a bias that predicts the classes belonging to the majority and frequently overlooks the classes belonging to the minority. As a direct consequence of this, the error rate rises for the groups that are underrepresented, whereas it falls for the classes that are overrepresented. We solved the problem of the unbalanced dataset by employing a method called random oversampling, which we used to both datasets..

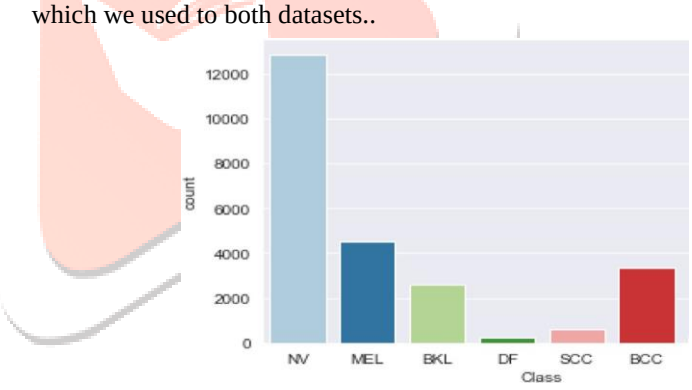


Fig.7 Image classification into 6 types.

5. Results and Discussion

5.1. Experiment

All of the tasks have been included into the python platform by our team. For the purpose of demonstrating that the suggested approaches are superior, we have applied them to the HAM10000 dataset. Before moving on to the work of feature extraction, we have finished the preprocessing step.

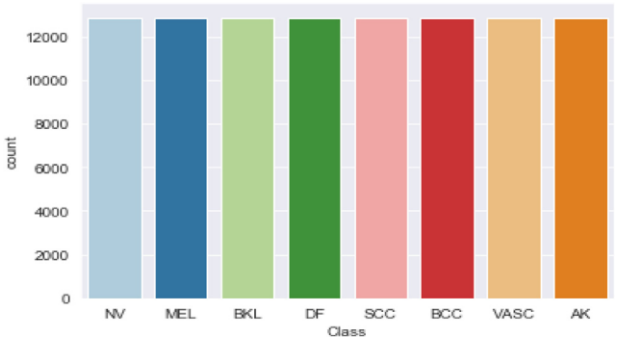


Fig. 8. Distribution after random Oversampling

In the first preprocessing step, we resize the input image into a size equal to 512 by 512 pixels. It reduces the amount of time required for processing while simultaneously increasing the system's overall efficiency. In the second phase, we got rid of the hairs by employing a digital hair removal (DHR) approach that was based on morphological filtering such the Black-Hat transformation and an image in painting algorithm. Last but not least, in order to get rid of the noise in the input photos, we used a Gaussian filter with a 7 x 7 kernel, which is also where the sigma value was determined. After Digital Hair Removal and Gaussian filtering, we calculated statistical measures like Root Mean Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similar Index Measure (SSIM), Spectral Angle Mapper (SAM), and Universal Image Quality Index (UIQ) for the images. Noise-to-signal ratio, peak signal-to-noise ratio, and structural similarity are examples. For the reader's convenience, the statistical measurements from the Fig. 4 photos' digital hair removal and Gaussian filtering are supplied in Tables 1. Following the completion of the preprocessing step, we extracted the foreground of the skin illnesses using a well-known segmentation approach known as k-means and automated ANN segmentation techniques. In contrast, the ANN technique performs significantly better than the k-means segmentation algorithm when it comes to identifying lesions in skin photos. Following the segmentation, there are still some noises present in the pictures; thus, we have utilized the Gaussian filtering approach once again in order to get rid of the sounds.

Table 1 : Statistical measures after applying Digital Hair Removal technique.

Image	RMSE	PSNR	SSIM	SAM	UIQ
ISIC_0000043	24.85	20.22	0.72	0.13	0.98
ISIC_0000103	10.26	27.91	0.85	0.06	0.99
ISIC_0000115	17.80	23.12	0.63	0.11	0.98
ISIC_0000218	12.40	26.26	0.82	0.07	0.99
ISIC_0000219	20.39	21.94	0.76	0.11	0.99

We initially demonstrated the classification performance of the proposed work for each of the three classification methods and compared them. Then, it shows that the model outperforms on the large dataset using the SVM algorithm (95% sensitivity) compared to the decision tree (93%) and even the KNN (94%), as shown in Fig. 11 and Table 2. It also explains the SVM algorithm's 95% sensitivity increase over the decision tree's 93%. However, the average accuracy and average F1 score values that can be produced by employing the SVM technique are both higher (95.13% and 94.88%). In addition, Table 2 demonstrates that the SVM classifier achieves superior results, ranking first among the three classifiers with an average accuracy of 95% and 97%, respectively, for both datasets. We have also analyzed our proposed model by utilizing the HAM10000 dataset, and we discovered that our model also works well when applied to that dataset, as can be shown in Figure 9 and Table 1, respectively. Table 7 shows that the Support Vector Machine classifier has a low amount of log losses for both datasets, which implies that it has a high degree of accuracy in its classification.

As can be seen in Figure 10, the overall performance that was achieved by the SVM classifier was superior to that which was accomplished by the other two classifiers.

Therefore, the value of the cell represents the proportion of accuracy of the forecast. In the confusion matrix, the diagonal cell displays the maximum level of predictions, which implies the lowest possible mistake rate for each category of skin conditions. In order to evaluate the efficacy of the suggested model, we first determined its accuracy, precision, recall, and f1-score through the use of the confusion matrix.

Table 2 Our balanced dataset models' performance metrics.

Dutta et al.	CNN (13 convolutional layers, 5 Max Pool, 3-FC layers and 1-output layer)	ISIC-2017	73.00	-
Yu et al.	Fully Convolutional ResidualNetwork (FCRN)-38, 50 ,101 layers	ISIC-2016	91.1	94.9
Proposed	**CNN (3 Convolutional blocks,Max Pool layers, 3-FC layers and 1-output layer)**	**HAM10000**	**82.28**	**91.2**

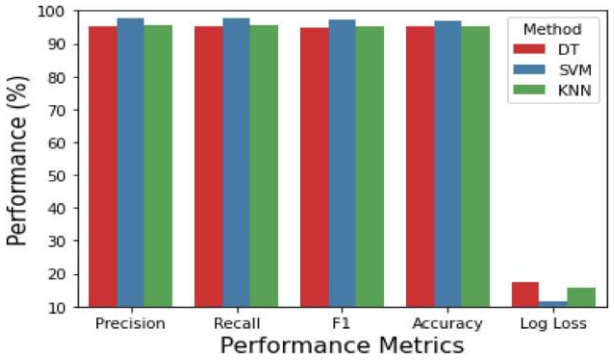


Fig. 9. HAM10000 dataset Precision, Recall, F1, Accuracy, and Log Loss comparison for three classifiers.

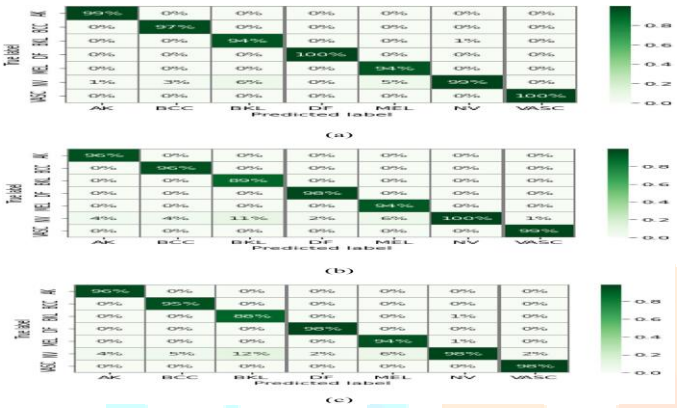


Fig. 10. The confusion matrix for the HAM10000 dataset of the proposed system employing the SVM, KNN, and Decision Tree classifiers to predict the eight classes' percentage values.

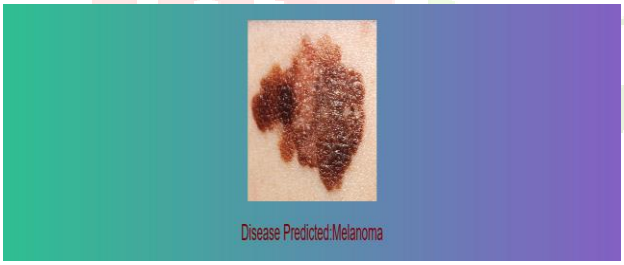


Fig 11: With proposed and Code Disease Predicted with 91.2% accuracy

6. Conclusion and Future Work

Diseases of the skin are today an issue on a worldwide scale. Skin illnesses may take numerous forms, and people from a wide variety of countries and areas suffer from them. The creation of new methods and procedures gives us more tools to use in the battle against these diseases. In the course of our investigation, we have carried out the task in numerous stages. We used a specialised approach of digital hair removal called Morphological Black-Hat Transformation to get rid of individual hairs and Gaussian Filter to blur the overall appearance of the photos. After that, we employed an artificial ANN segmentation to detect the skin lesion. This segmentation method was successful in locating the diseased region and segmenting it in a precise manner. In the end, we classified the different types of skin diseases by extracting the GLCM and a few statistical features and then applying those features to the SVM, KNN, and DT classifiers. We utilized two benchmark datasets that are available to the public: HAM10000. As a result of the fact that these datasets are unbalanced in some way, we decided to undertake data balancing by employing the random over sampling approach. When we used SVM, KNN, and DT classifiers, respectively, on the HAM10000 dataset, we were able to get an average accuracy of 96%, 82.28%,79.28% and 96.71% respectively. It shows that when we use the HAM10000 dataset, our model performs significantly better. In addition to this, we found that our model works exceptionally well when applied to balanced data. When compared to other approaches that are considered to be state-of-the-art for the categorization of skin diseases, our model performs much better.

This model may also be utilized for the purpose of completing additional skin disease categorization tasks. Nevertheless, there is room for advancement in terms of the performance of the categorization. We made use of an artificial segmentation approach, which might at times result in inaccurate detection of the skin lesion. As a consequence of this, it contributes to incorrect categorization, which is one of the limitations of our research. In the future, research will concentrate on the diagnosis of skin diseases in real time. This will be made possible with the assistance of methods of segmentation and classification that are more effective, such as techniques of ensemble learning and deep learning. In addition to this, we think that it will improve the efficiency as well as the accuracy of the algorithms that are used in picture classification and object identification systems. We have high hopes that it will be of assistance to patients for the early diagnosis of disorders in order to maintain the health of their skin.

References

[1] Anatomy of the skin, Stanford children's health, 2021, [Online]. Available: https://www.stanfordchildrens.org/en/topic/default?id=anatomy-of-the-skin-85-P01336 (Accessed 11 June 2021).

[2] M.W. Greaves, Skin disease, Britannica, 29, 2020, [Online]. Available: https://www.britannica.com/science/human-skin-disease. (Accessed 11 June 2021).

[3] Bangladesh: Skin disease, 2018, [Online]. Available: https://www.worldlifeexpectancy.com/bangladesh-skin-disease. (Accessed 9 July 2021).

[4] Cancer, world health organization, 21, 2021, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. (Accessed 20 March 2020).

[5] Md. Al Mamun, Mohammad Shorif Uddin, A

survey on a skin disease detection system, Int. J. Healthc. Inform. Syst. Inform. 16 (4) (2021) 1–17.

[6] C.N. Vasconcelos, B.N. Vasconcelos, Experiments using deep learning for dermoscopy image analysis, Pattern Recognit. Lett. 139 (2020) 95–103.

[7] U.-O. Dorj, K.K. Lee, J.Y. Choi and. M. Lee, The skin cancer classification using deep convolutional neural network, Multimedia Tools Appl. 77 (2018) 9909–9924.

[8] M. Taufiq, N. Hameed, A. Anjum, F. Hameed, m-Skin Doctor: A Mobile Enabled System for Early Melanoma Skin Cancer Detection Using Support Vector Machine, in: eHealth 360°. International Summit on eHealth, 2017, pp. 468–475.

[9] Jagdis etal., J.A.D.L. Cruz-Vargas, M.E.R. Camacho, Advance study of skin diseases detection using image processing methods, Nat. Volatiles Essent. Oils J. 9 (1) (2022) 997–1007.

[10] Z. Naeem, G. Zia, Z. Bukhari, A healthcare model to predict skin cancer using deep extreme machine, J. NCBAE 1 (2) (2022) 23–30.

[11] S.K. Bandyopadhyay, P. Bose, A. Bhaumik, S. Poddar, Machine learning and deep learning integration for skin diseases prediction, Int. J. Eng. Trends Technol. 70 (2) (2022) 11–18.

[12] A. Kalaivani, S. Karpagavalli, Detection and classification of skin diseases with ensembles of deep learning networks in medical imaging, Int. J. Health Sci. 6 (S1) (2022) 13624–13637.

[13] S.A. AlDera, M.T.B. Othman, A model for classification and diagnosis of skin disease using machine learning and image processing techniques, Int. J. Adv. Comput. Sci. Appl. 13 (5) (2022).

[14] P.R. Kshirsagar, H. Manoharan, S. Shitharth, A.M. Alshareef, N. Albishry, P.K. Balachandran, Deep learning approaches for prognosis of automated skin disease, Life 2022 12 (426) (2022).

[15] M.Q. Hatem, Skin lesion classification system using a Knearest neighbor al- gorithm, in: Visual Computing for Industry, Biomedicine, and Art. Vol. 5, (7) 2022.

[16] K.S. A, M.S. B, Melanoma disease detection and classification using deep learning, Int. J. Res. Appl. Sci. Eng. Technol. 10 (7) (2022).

[17] P. Yao, Single model deep learning on imbalanced small datasets for skin lesion classification, IEEE Trans. Med. Imaging 41 (5) (2022) 1242–1254.

[18] K. Padmavathi, H. Neelam, M.P.K. Reddy, P. Yadlapalli, K.S. Veerella, K. Pam- pari, Melanoma detection using deep learning, in: 2022 International Conference on Computer Communication and Informatics, ICCCI, 2022.

[19] M. Maduranga, D. Nandasena, Mobile-based skin disease diagnosis system using convolutional neural networks (CNN), I.J. Image Graphics Signal Process. 3 (2022) 47–57.

[20] A. Jain, A.C.S. Rao, P.K. Jain, A. Abraham, Multi-type skin diseases classification using OP-DNN based feature extraction approach, Multimedia Tools Appl. 81 (2022) 6451–6476.