EMERGENT: JOURNAL OF EDUCATIONAL DISCOVERIES AND LIFELONG LEARNING (EJEDL)
ISSN: 2776-0995 Volume 3, Issue 4, April, 2022

SKIN CANCER IMAGES CLASSIFICATION USING NAÏVE BAYES ALGORITHM

Ohood Fahdil Alwan
College of Al Muqdad, University of Diyala, Diyala, Iraq
oh85ood@gmail.com

Abstract

It might be well known that skin cancer is one of the foremost perilous sorts of cancer, essentially, there are two sorts of skin cancer named malignance and non-malignance. In current paper the system is designed for the exposure and classification of skin cancer with high precision and exactness by using Naïve Bayes which capable to diagnoses different sorts of cancer in a human skin. It objects at utilizing more significant information to develop the skin cancer and assistance physicians in the clinical to diagnosis and exact detection of disease. The designed system has an important role in to avoid errors during the while identification and sorting of cancer. This is encouraged by probable execution enhancement in the overall automatic and giving confidence in decision-making and hasty detection of skin cancer. This technology is of extraordinary and financial significance to doctors. The system is separated into two sorts (system with pre-processing and another system without pre-processing). The former system contains the following stages: image acquisition, preprocessing, and classification, whereas the last system contains of image gaining and classification. The results from this paper show that the model using (NB) without any pretreatment has an average level of accuracy 70.15%, however with preprocessing has an accuracy of 69.69%. The diminished precision returns to the reason that the skin pictures that are taken to the skin are as well near and they don't require any handling.

Keyword: Skin Cancer, Naïve Bayes Algorithm, data Pre-processing.

Introduction

Skin is a vital tissue that safeguards the whole outer humans' body, function to a defensive wall against pathogens of the location. Its presence in the outer part, the skin is set to sickness. It is common as "skin cancer". The skin cancer is abnormal in skin cells produced by changes in cell (DNA). Melanoma cancer is the riskiest type of the skin cancer. The skin pigment cells that products melanin. Melanoma is mostly colored by brown or black, because cells are still able to form melanin [World Health Organization. 2018].

More than 5,400 people worldwide die every month from malignant skin cancer. According to the statistically evaluated that the numbers of new cases of melanoma

cancer are diagnosed in 2020, will increment nearly to 2 %. It is expected that the number of deaths by skin cancer will decrease to 5.3 % in 2020.

Out of these cases, 60,190 are men and 40,160 are women. According to the statistics of the past decade, (2010-2020), the number of novel diagnostic melanoma cases were diagnosed annual have enlarged by 47 % (Cancer Facts and Figures 2020).

To process the problems caused by limited access to specialists, especially in evolving countries, there has been a important amount of research focused on developing automated image analysis systems based on dermatoscopy images that detect skin diseases.

There are numerous new publications, reviewing various used techniques like this one (Singh et al., 2020) as well as dermoscopy papers, which develop diagnostic standards for initial melanoma discovery (Mishra and Celebi., 2016).

Automated image analysis varied widely, but in the computer vision curriculum, it is considered restricted as it uses groups of representations of optical features that have a low level such as color, description, edge, and the determination of colored- melanin, etc. One of the algorithms that included fractionation of the lesion, such as an support vector machines (SVM) and k-nearest neighbor (KNN) in which image processing is based on rules (Garnavi., Aldeen., and Bailey., 2012) (Celebi et al., 2007).

The image classification and the detection of the skin cancer have been suggested by lessons proposed. There exists a mortification of research papers. A comprehensive survey of these approaches is available in Ref. (Masood, and Ali., 2013), (Kittler et al., 2002), (Erkol et al., 2005). Every one of these papers used available state-of-art styles and improvement of claim performance. The most known approaches used for classifications of image differ from application of algorithms of decision tree (Zorman et al., 1997), (Friedl, and Brodley., 1997) and Bayesian classifiers (Larsen, 2005) (Ruiz et al., 2011) and supporting of vector machines (Gilmore., Hofmann., and Soyer., 2010), and diversity of Artificial- based Intelligence approaches (Silver et al., 2016) (Codella et al., 2015).

In this study, a system was designed to identify and classify melanoma, the researcher used (Naïve Bayes) as a tool to classify the skin cancer according to their appropriate categories, which are (B) the Benign, (M) Malignant. It aims to use more meaningful data to improve skin cancer detection and to assist physicians in clinical diagnosis for extra accurate disease detection. This system assists clinicians in avoiding mistakes while classifying and diagnosing cancer kinds. This is motivated by the possibility for improved general automatic performance, which provides consistency in decision-making and allows for the early diagnosis of skin cancer. This technology is extremely valuable to clinicians in terms of both time and money.

Literature Review

Machine learning is the arena of study that provides the computers the ability to learn with no need for complicated systems, the classification of the skin cancer was discussed using naive bayes classifier with shearlet transformation factors with three coefficients. (Mohan Kumar, 2019), treated Melanoma images for rank feature then applied naive bayes for classification. By using the PH2 dataset, which contains 100 dermoscopic RGB images with melanocytic lesions with resolution is768x560 pixels. Some proposed models depend on advantage of concepts of the naive bayes probability classifier (Park, 2016) . In case of proposed a model, it explains how the naive Bayes Classifier could classify image of the skin cancer and can be formed as the maximum posteriori of decision-making rule in order to reduce the training time for the algorithm. Total images of dataset 800 images divided in four class and each class contains 200 images. The proposed classifier reached to an accuracy of 77.2 %. Using the improvement of image pre-processing in noise removal, the skin disease images are obtained (Lowd,and Domingos 2005). The images are then segmented using K-means by gathering process. During the feature extraction phase, the segmented images are introduced, and extracted images are categorized using **a** classification technique such as Naive Bayes.. This could make out a good result.

Classification

There are two types of classification methods: supervised and unsupervised learning. Clustering is a standout strategy in the field of unsupervised learning. The realizing algorithm of a neural network it is possible to be supervised or unsupervised. When the desired output is already known, however, supervised learning techniques are appropriate. Though, unverified kind of learning algorithms is employed if not objective output is available.

Naïve Bayes

"Naive Bayes" assorted is raised on "Bayes' Theorem" of possibility. (Hart et al., 2000). In Bayes' theorem, the subjunctive possibilities that an occurrence x mentioning to a class k possible to finding specific incidents in each kind of the conditional probabilities and the unconditional possibility of the incident in every kind. That is, for given data, $x \in X$, and C types, where X signifies a random mutable, the conditional probability for incident drives to a category can be used to calculated the equation:

$$P(x) = P(c_k)\frac{P(x)}{P(x)} \qquad (1)$$

Equation (1) illustrates that the computing of $P(x)$ is a form of classification problem since it refers to the probability that the assumed data x belongs to class k and the researcher can choose the best category by choosing the type , which has the highest probability of all possible types. C, which the classification error can be minimize. To

achieve this, we need to guesstimate $P(x|c_k)$ and agree any particular value of vector(x) conditional on(c_k) is statistically independent of every dimension can be printed as:

$$P(X|C_k) = \coprod_{i=0}^{n} \quad P(X|C_{jk}) \qquad (2)$$

where x is a (n- dimensional) vector of data $x = (x_1, x_2, \cdots, x_n)$.

The Naive Bayes classifier is created on eq. (2) and adopts that all feature be statistically independent [20]. This supposition results in simpler computation loss and effective image processing. By joining eq. (1) and eq. (2), the Naive Bayes classifier can be brief as equation:

$$k = argmax_k P(C_k) \coprod_{i=0}^{n} \quad P(X_i|C_k) \qquad (3)$$

where the denominator $P(x)$ is gone since the value is the similar for all class.

The Naive Bayes classifier is regularly denoted as the highest A posteriori (MAP) decision rule. Notes that the assumption of statistically independence in every feature from time to time does not hold confident cases and causes issues in some practical situations (Jiang et al., 2005]. However, even when the assumption does not hold, Naive Bayes assumptions yield an optimal classifier, several applications and experimental studies display that training structures utilizing on the(MAP) decision instruction with the Naïve Bayes.

Proposed System

In this study the researcher proposes a technique that focuses on detecting and classify of the skin cancer from image. The input to the system is an image, and the output is a classification of skin cancer. Figure (1) illustrates the main aspects of the suggested method.

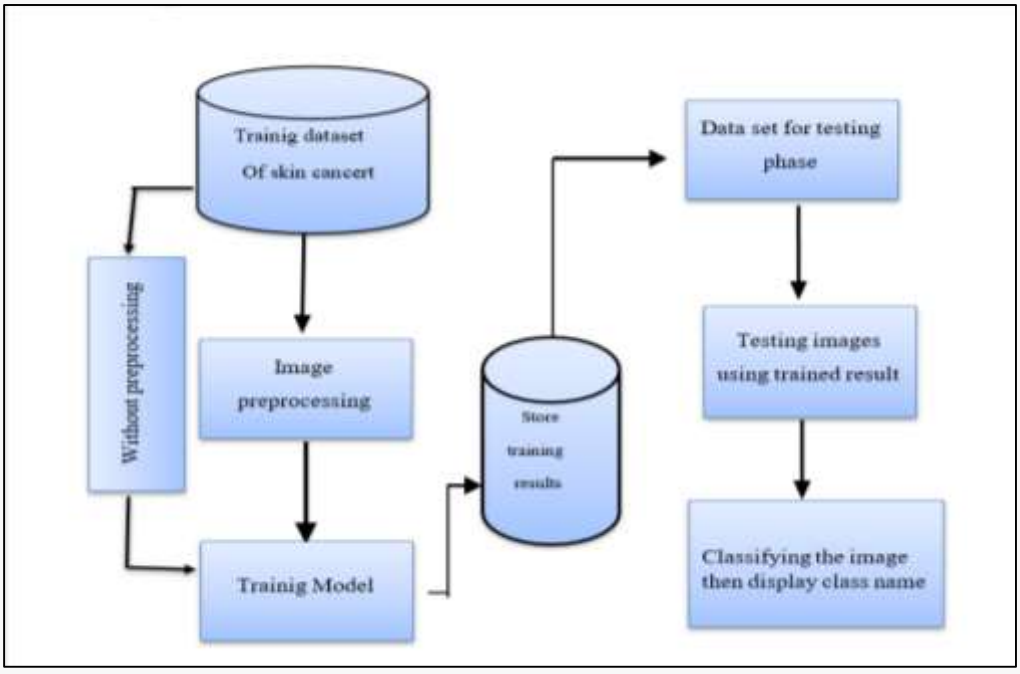

Figure 1: Block Diagram of the of Proposed Model

The proposed model has several stages and each stage contains several steps that work together to achieve the system goals. The system is divided into two types (one with pre -processing and the other without pre - processing). First one contains from: image gaining, preprocessing and tabulation, whereas the second one contains of image acquisition and classification.

Image dataset

The datasets that are used in proposed system has 3297 some for training, and others for testing. In fact, there are more than (1497) appearance cases of malignant skin caser type, and (1800) images cases for benign carcinoma. The entire whole dataset material was collected from the ISIC (International Skin Image Collaboration) Archive. All the images have been resized to lower resolution (224*224*3) RGB. The images of skin cancer have (24-bit) RGB color space where each (8 bits for each channel). Table (1) shows the distribution of skin cancer images. And Figure (2) presents the samples of these types of skin cancer.

Table (1) Distribution of Number Skin Cancer Images Dataset

Dataset	Training	Testing
Benign	1440	360
Malignant	1197	300

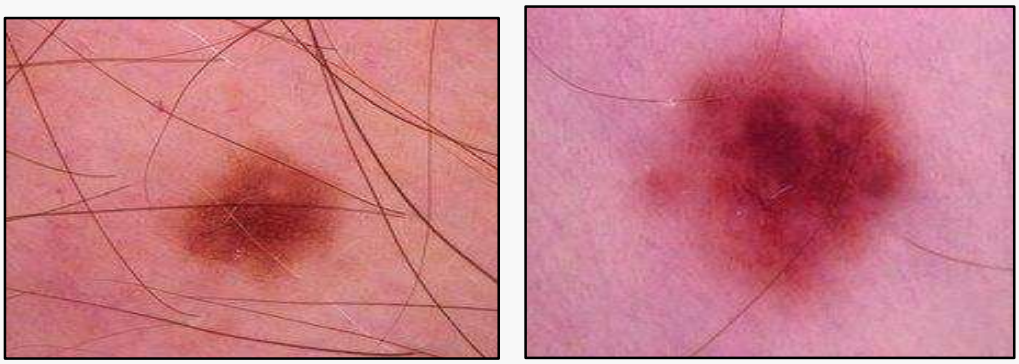

Figure (2): Samples of Different skin cancer Images

Image Preprocessing Stage

In preprocessing image step the proposed system consists steps, they are:

Normalization for image

Normalization for Image is a method in which change is occurred in the range of pixel intensity values in order to make the pictures more familiar or normal to the minds. As the images have already been resized to 224*224, there is no need to resize them,

then normalize in rang [0,1]. All the values of the images are extracted by dividing them RGB values of images by 255.

Removal of hair by image closing operation and median filter

One of the most common artifacts is hair that must be removed in skin pictures. There are many algorithms and methods in the literature for removing hair when it is not shaved after the image acquisition process. Hair algorithm depends on two basic stages

1. Firstly, apply simple closing morphological process with a (disk-shaped) structured element. Then the hair segments are thin structures., these morphological procedures are applied.
2. Next, a hair mask is reserved by applying a universal automatically threshold on the image strength information. An average mean of the neighbor's pixels exchanged by each hair pixel from the caused mask is. (using median filter after closing operation).

Algorithm (1): Hair Remove by image closing followed by median filter
Input: RGB Image with size 224*224 L= 223, n=8 **Output**: RGB Skin cancer image
Begin **Step (1):** Input an image **Step (2):** The two-dimension window of size 3×3 is selected from image **Step (3):** For i = 1 To L-1 **Step (4):** For j = 1 To L-1 **Step (5):** Apply image closing for each layer R layer Apply median filter on image with size (3*3) G layer Apply median filter on image with size (3*3) B layer Apply median filter on image with size (3*3) **Step (6):** Test the removal of the hair load the image convert RGB Repeat from step 2 to 3 **Step (7)**: Display the RGB image END

Training step our classification algorithm was applied by selecting classes to be used for training and testing. The main purpose of the training step is computing prior

probability of each class $p(c_i)$ and likelihood $p(x|c_i)$. And to build the feature vector of class C_i where i is the number of classes. Each class C_i has a set of training images$\{p_1 p_2 p_3 \dots p_n\}$ where n is the features number of in each class.

Prediction step: Any invisible test data (X), the method calculates the posterior of probability $p(c_i|x)$ of that sample referring to each class. The method then classifies the test images based on the maximum posterior probability. Algorithm (2) summarizes the step of (NB) algorithm -without Preprocessing

Algorithm (2): Naive Bayes algorithm -without Preprocessing
Input: Training data set TD, **Output:** Class Name
Begin **Step (1):** Read training Dataset DS**.** **Training phase (80) % of datasets was taken for training.** **Step (2):** (Total) = all examples in training dataset. C_j refer to class in training DS. **Step(3) :** Calculate the Probability for each class. **P(C_j)** = frequency (C_j) / total. **Step (4):** Calculate the mean (to feature) (μ) and standard deviation (to feature) (σ) values for each feature in each class in training dataset and store the result **Testing phase (20) % of datasets was taken for testing** **Step (5):** X is tested example in the testing DS **Step (6):** Calculate the probability of density function (pdf) of X at C_j, for features of (X) exists in S, p (X_i | C_j) by apply the eq. (2). **Step (7):** Calculate conditional probability of(X) at C_j for values product from step (6), by apply equation. $P(X|C_i) = \coprod_{j=1}^{n} P(f_j|C_i)$ **Step (8):** Calculate posterior probability of(X), p (C_j|X) that denote probability of example at C_j by equation: $P(C_i|X) = P(C_j)p(C_j)$ --- probability of feature at (C_j). **Step (9):** Select class label to class (X) by choice maximization p (C_j|X). **Step (10):** Return Name of Class. **END**

The same steps are repeated for the proposed Naïve Bayes classifier but with preprocessing. The steps are taken to clear these steps for naïve bayes. The algorithm is listed in the algorithm (3).

Algorithm (3): Naive Bayes algorithm -with Preprocessing
Input: Training data set **Output: Name of Class**

Begin
Step (1): Read the training Dataset DS.
Step (2): Apply Preprocessing on images as algorithm (3.3).
Training phase (80) % of datasets was taken for training.
Step (3): Repeat the same steps (2 to 4) in algorithm (3.6) //For the training stage
Testing phase (20) % of datasets was taken for testing
Step (4): Repeat the same steps (5 to 10) in algorithm (3.6) //For the testing **stage.**
END

Results and Discussion

The methods in this paper are applied as follows: NB to explore the best performance model. To achieve the research objective, the model is divided into two parts. The first part is without preprocessing for dataset and the results as show in figure (3) and table (2) which indicate the accuracy of classification of this model.

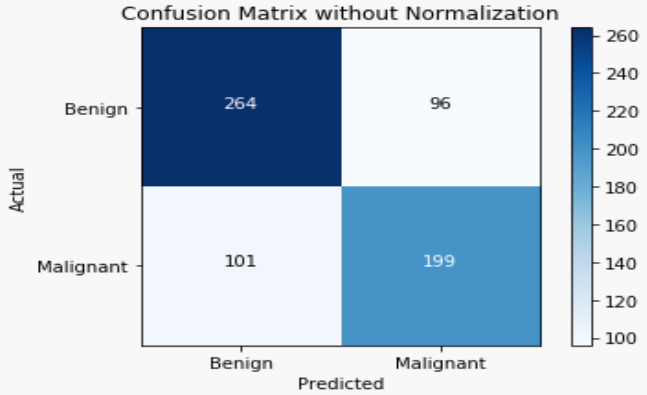

Figure (3) The Confusion Matrix for Naïve Bayes without preprocessing

Table (2) Naïve Bayes Accuracy without preprocessing

Accuracy	70.15%
Sensitivity	66.33%
Specificity	73.33%
Precision	67.45%

Figure (3) and Table (2) illustrates the classification rate of the first type of proposed system. In classification system for skin cancer images that presented by applying

(NB) classifier is put forward. The algorithm (NB) does not require great deal of training extreme of data compared to neural networks and their artificial structure. This makes it easier for the classifier to make a correct decision and with limited calculations. To reduce the training time, it is better for the proposed classifier to utilize Naive Bayes classifier instead of the traditional classifiers as producing correct results in classification by approving the advantage of Naive Bayes algorithm and concepts of probability to the classifier.

The second part of our model is evaluated by using the same dataset, the values of the accuracy, sensitivity, specificity, and precision of model (with preprocessing) illustrated in figure (4) and table (3). The classification accuracy of the second proposed model with preprocessing and is 69.69%, and the total time is 540 seconds.

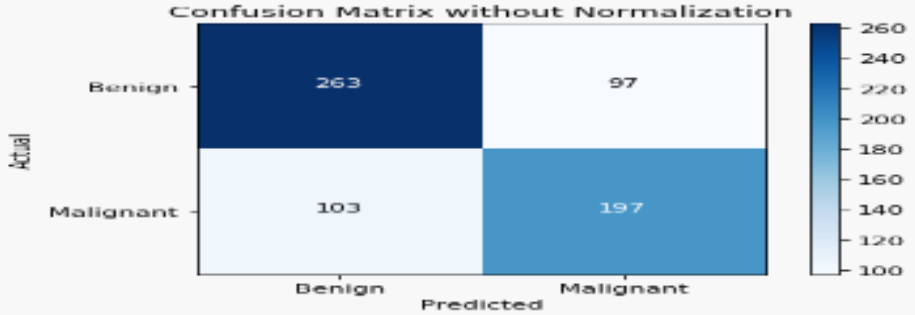

Figure (4) The Confusion Matrix for Naïve Bayes with preprocessing

Table (3) Naïve Bayes Accuracy with pre processing

Accuracy	69.69%
Sensitivity	65.66%
Specificity	73.05%
Precision	67.00%

From the tables above, the convergence is clearly noticed in the rate of classification of accuracy between our model (with or without preprocessing) as the preprocessing is not affected by the improvement of classification accuracy. As for the training time spent in our model in the classification of skin cancer images, a naive Bayes algorithm (without preprocessing) takes less time than same algorithm but (with preprocessing).

Table (4) difference between Naïve Bayes with and without processing

Type	Accuracy	Time
Naïve Bayes with Pre processing	69.69%	597 seconds
Naïve Bayes without Pre processing	70.15%	540 seconds

Conclusion and Future Work

Skin cancer is classified in both benign and malignant types using a model such as NB and its proposed system with or without pre-processing. As the data is initially trained to obtain a trained mode. After that this model is tested along with the testing of the

data. Finally, the results show the prediction in the form of the probability of each type of the skin cancer.

The results show that the data, which are utilized in the training, affect the level of accuracy in the classification of skin images of cancer patients. It becomes clear that the model without pre-processing data has a higher accuracy in terms of classification than the model when using pre-data processing. The reason is that the training data set does not need to be pre-processed. However, when the modal of preprocessing is applied on image, it is likely to lose some of its information.

In case of doing similar potential study like the current study, it is suggested to use the other types of dermal cancer data set, as well as it is recommended to approach other methods of pre-processing the data and then attempt to make comparison with findings of the current study.

References

1. Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. Remote sensing of environment, 61(3), 399-409.
2. Zorman, M., Štiglic, M. M., Kokol, P., & Malčić, I. (1997). The limitations of decision trees and automatic learning in real world medical decision making. Journal of Medical Systems, 21(6), 403-415
3. Hart, P. E., Stork, D. G., & Duda, R. O. (2000). Pattern classification. Hoboken: Wiley.
4. Kittler, H., Pehamberger, H., Wolff, K., & Binder, M. J. T. I. O. (2002). Diagnostic accuracy of dermoscopy. The lancet oncology, 3(3), 159-165.
5. Erkol, B., Moss, R. H., Joe Stanley, R., Stoecker, W. V., & Hvatum, E. (2005). Automatic lesion boundary detection in dermoscopy images using gradient vector flow snakes. Skin Research and Technology, 11(1), 17-26.
6. Jiang, L., Zhang, H., Cai, Z., & Su, J. (2005, April). Learning tree augmented naive bayes for ranking. In International Conference on Database Systems for Advanced Applications (pp. 688-698). Springer, Berlin, Heidelberg.
7. Larsen, K. (2005). Generalized naive Bayes classifiers. ACM SIGKDD Explorations Newsletter, 7(1), 76-81.
8. Lowd, D., & Domingos, P. (2005, August). Naive Bayes models for probability estimation. In Proceedings of the 22nd international conference on Machine learning (pp. 529-536).
9. Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., & Moss, R. H. (2007). A methodological approach to the classification of dermoscopy images. Computerized Medical imaging and graphics, 31(6), 362-373.
10. Gilmore, S., Hofmann-Wellenhof, R., & Soyer, H. P. (2010). A support vector machine for decision support in melanoma recognition. Experimental dermatology, 19(9), 830-835.

11. Ruiz, D., Berenguer, V., Soriano, A., & Sánchez, B. (2011). A decision support system for the diagnosis of melanoma: A comparative approach. Expert Systems with Applications, 38(12), 15217-15223.
12. Garnavi, R., Aldeen, M., & Bailey, J. (2012). Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis. IEEE transactions on information technology in biomedicine, 16(6), 1239-1252.
13. Masood, A., & Ali Al-Jumaily, A. (2013). Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. International journal of biomedical imaging, 2013.
14. Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., & Smith, J. R. (2015, October). Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In International workshop on machine learning in medical imaging (pp. 118-126). Springer, Cham.
15. Mishra, N. K., & Celebi, M. E. (2016). An overview of melanoma detection in dermoscopy images using image processing and machine learning. arXiv preprint arXiv:1601.07843.
16. Park, D. C.(2016).” Image Classification Using Naïve Bayes Classifier” International Journal of Computer Science and Electronics Engineering (IJCSEE) Volume 4, no.3 pp 2320–4028 .
17. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), 484-489.
18. World Health Organization. 2018. Skin Cancers. Available: http://www.who.int/uv/faq/skincancer/en/index1.html.
19. Mohan, K., Ram, K., Gopalakrishnan, K.,” Skin Cancer Diagnostic using Machine Learning Techniques – Shear let Transform and Naïve Bayes Classifier” International Journal of Engineering and Advanced Technology (IJEAT) Vol.9 no.2, pp :2249 –8958. 2019.
20. Cancer Facts and Figures 2020.(ACS) https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures.
21. ISIC Database (International Skin Image Collaboration)
22. Singh, R. K., Gorantla, R., Allada, S. G., & Pratap, N. (2020). Ski Net: A Deep Learning Solution for Skin Lesion Diagnosis with Uncertainty Estimation and Explain ability. arXiv preprint arXiv:2012.15049.