I FOR YOU: An Intelligent Chatbot System for Skin Cancer Detection Using CNN

Sneha K
BCA(AI)
Rathinam College Of Arts And Science
Coimbatore,India
kannansneha622@gmail.com

Vivek.P
BCA(AI)
Rathinam College Of Arts And Science
Coimbatore,India
pvsvivek21@gmail.com

Vasanth.M
BCA(AI)
Rathinam College Of Arts And Science
Coimbatore,India
Itsmevasanth65@gmail.com

Abstract

Skin cancer is one of the most common cancer types and leads to hundreds of thousands of yearly deaths worldwide. For the sake of raising people's serious awareness of skin cancer and providing a convenient way to diagnose skin cancer, this study designs a chatbot to help users figure out the true situation of their skin. In this field, many current studies still cannot achieve high accuracy for skin cancer detection.

The probability of recovery significantly improves with timely detection. Recent advances in deep learning artificial intelligence have led to significant progress in image-based diagnosis.To increase the accuracy, deep learning method convolution neural network (CNN) was used to detect the seven types of tumors, using the HAM1000 dataset. This dataset comprises 10,015 dermatoscopic images The photos were augmented, normalized, and resized during the preprocessing step. Skin lesion photos could be classified using a CNN method based on an aggregate of results obtained after many repetitions.In this model, a chatbot named "I FOR YOU" is created based of NLP and SpaCy.The cmodel created can diagnose seven types of diseases.finally, the results indicated that the model achieves higher accuracy and makes the interaction with the user becomes true.

Keywords: *Convolutional Neural Network, Chatbot, Skin Cancer Detection, NLP, SpaCy, Opencv*

I. Introduction

Skin cancer is a prevalent disease worldwide and can be categorized into two groups: nonmelanoma and melanoma. Melanoma skin cancer is the more severe type, accounting for 75% of all skin cancer-related deaths.In the U.S., statistics showed that nonmelanoma skin cancers have constituted one third of all cancers [1]. . For melanoma skin cancer, study has shown that the rate has been steadily increasing by about 3%to 7% annually. The rates have become so high that about 1% of the U.S. population born in 1993 will develop malignant melanoma skin cancer [1]. Furthermore, the areas most affected by skin cancer around the globe are shown in ***Figure 1***, with North America accounting for about half of the total.

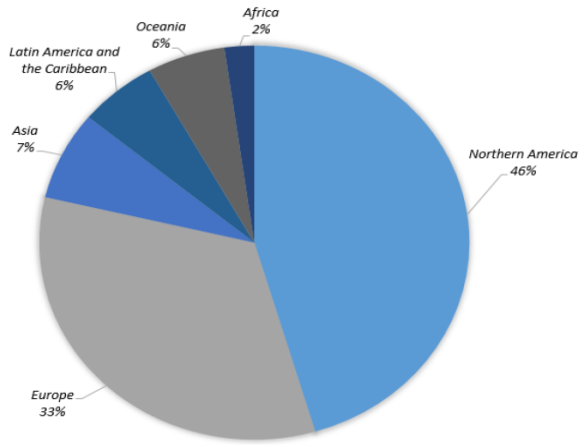

Figure 1. Skin cancer cases globally (22 March 2022) [1].

Early detection of skin cancer is crucial in preventing related deaths. The conventional diagnostic process involved physicians questioning patients about their personal information such as family medical history and gender to identify those at high risk. Next, the examiner would inspect every part of the patient's skin carefully with the naked eye or using microscopy to detect potential skin cancer based on appearance Asymmetry, Border, Color, Diameter (ABCD) checklist is often used to assist detection [2].

The diagnostic procedures for skin cancer can be lengthy and their accuracy may vary depending on the proficiency of the physician. Consequently, there is a need for a more efficient and dependable approach. One option being researched is the utilization of computers to analyze photos of suspected spots and detect skin cancer. While computer technology has been employed to assist in diagnosing skin cancer, there are two major shortcomings in the current algorithms. The first is related to the selection of models. For instance, in a study [3], the authors employed artificial neural

networks (ANN), fuzzy rule-based systems, or adaptive fuzzy inference neural networks (AFINN) as their classifiers. Note that AFINN would adjust weights using the Backpropogation (BP) algorithm. Similarily, in [6], the authors utilize BP in addition to ANN as their classifiers.However the traditional neural network has its limitations.The current algorithm faces another problem which is related to the use of outdated datasets to train the models.

Back in those days, the datasets used had only two skin classifications - benign lesions and melanoma - with no additional details or subclasses. In contrast, the ISIC 2018 dataset, which is more recent, contains over 10,000 images.
The current algorithm and datasets have issues that need to be addressed through the implementation of a better structure. Given the success of convolutional neural network (CNN) in
various fields, including medicine (such as identifying lung cancer from CT images [9]) and industry (such as enabling automatic driving [5]), this study has chosen to utilize CNN for further image processing. The proposed model appears to successfully address the aforementioned problems.Also, to settle the problem of the too small dataset, a new dataset of skin cancer, named concentrated dataset, a new dataset of skin cancer, named HAM1000, was used in this study.

II. Methods

1. Related Work

The accuracy of computer-aided techniques was evaluated by experts who examined the strength of the supporting facts [4]. They consulted ScienceDirect, SpringerLink, and IEEE databases to analyze skin lesion segmentation and classification approaches and identified significant limitations. In [12], an improved technique for diagnosing melanoma skin cancer was presented. The technique used an implantation manifold with nonlinear embeddings to generate synthetic views of melanoma and employed data augmentation to create a new collection of skin melanoma datasets. The SqueezeNet deep learning model was trained using the enhanced images, and the experiments demonstrated a significant improvement in melanoma identification accuracy (92.18). In [13], the VGG-SegNet algorithm was proposed for extracting a skin melanoma (SM) region from a digital dermatoscopy image. The extracted segmented SM was compared with the ground truth (GT) to establish essential performance parameters. The proposed scheme was evaluated and verified using the standard ISIC2016 database.

To categorize skin melanoma at an early stage, researchers offered a deep-learning-based methodology, including a region-based convolutional neural network (RCNN) and fuzzy k-means clustering (FKM) [11]. The suggested technique was put to the test using a variety of clinical photos in order to aid dermatologists in the early detection of this life-threatening condition. The ISIC-2017, PH2, and ISBI-2016 datasets were used to assess the provided methodology's effectiveness. The findings revealed that it outperformed current state-of-the-art methodologies with an average accuracy of 95.40%, 93.1%, and 95.6%.

DL models such as convolutional neural networks (CNNs) have proven themselves superior to more traditional methods in various fields, especially image and feature recognition [14]. Moreover, they have been effectively applied in the medical profession, with phenomenal results and outstanding performance in a variety of challenging situations. Doctors and professionals now have access to a variety of DL-based medical imaging systems to aid in cancer prognosis, treatment, and follow-up assessments.The Lesion-classifier, relying on pixel-by-pixel classification findings, was presented to categorize skin lesions into melanoma and non-melanoma cases. Skin lesion datasets ISBI2017 and PH2 were used in the investigation to verify efficacy. The experiments showed that the suggested technique had an accuracy rate of 95% on the ISIC 2017 and PH2 datasets [15].

III. Proposed System

3.1 Data Preparation:

The limited size and lack of diversity of the available dataset of dermatoscopic images are obstructing the training of neural networks for the automated diagnosis of pigmented skin lesions. This work uses the dataset called HAM1000 (Human Against Machine), which is comprised of 10,015 dermatoscopic images collected from various populations and acquired and stored by different modalities. The images are divided into seven categories:

Class	Number Of Images
Actinic keratosis and intraepithelial carcinoma (AKIEC)	327
Basal cell carcinoma (BCC)	514
Benign keratosis-like lesions (BLS)	1099
Dermatofibroma (DF)	115
Melanoma (Mel),	1113
Melanocytic nevi (NV),	6705
Vascular lesions (VASC)	142

Some images from the dataset:

Figure 1

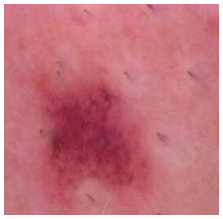

Figure 2

Figure 3

Figure 4

3.2 Convolutional Neural Network

Then structure of this study is the CNN model,which is widely used model for image processing.A CNN is a type of deep learning technique that is specifically designed for analyzing visual image-based data. Like other artificial neural networks, CNNs are trained to learn the features of the training data and use this knowledge to distinguish between different classes in the test data, using a combination of feedforward and backpropagation. Although CNNs typically require more computational resources than basic machine learning techniques, they also tend to deliver superior performance. In a typical CNN, there are three types of layers - convolutional layers, pooling layers, and fully connected layers - arranged in a specific order. The complexity of the network and its ability to identify patterns increases with the number of layers.

The design of the visual system inspired CNN architecture, which mirrors the connection pattern of the human brain neurons. The basic architecture of the model shown in Fig.8. CNN segments images using filters to make them easier to process without losing vital details. A convolution layer is created by convolving the image as input with a filter to extract low-level characteristics like colors and sharp edges. To lower the processing power required for data processing, a pooling (max-pooling or pooling layer) layer reduces the size of this layer with extracted characteristics. One layer of CNN is made up of these two layers. A chosen number of layers can be put into the training model and collect minimal stage characteristics based on the picture's complexity. The output is then flattened to feed it to a conventional complete layer for image classification. The system is then enhanced by an entire layer that learns highlevel characteristics via nonlinear mixtures processed by the convolution layer.

After that, the image is flattened into a column vector and fed into a feed-forward neural network. Every cycle of the training process is then subjected to backpropagation.

Through a sequence of epochs, the model learns to differentiate between dominant and limited characteristics of images to classify them.

3.3. Tensorflow

TensorFlow isn't as simple as Keras, but it has more features and capabilities for developing machine learning models. It's a Python open-source software library. Data flow graphs for machine learning models are created with it. TensorFlow models can be run on any platform, including locally on a personal computer, the cloud, Android, and iOS. It's simpler to use than CNTK because it allows you to focus entirely on the model's logic. It also allows the developer to observe the neural network that has been created.

3.4. Keras Library

Keras is a user-friendly high-level API that is exceedingly easy to use. It's a Python-based deep learning package that's free source. It can be used as the frontend for other Deep Learning packages, such as TensorFlow.

The construction of the CNN model will be aided by combining TensorFlow and Keras. Deep Learning capabilities that are more advanced the backend of the build process will use TensorFlow. At the same time, the frontend, Keras, will be utilized because of its high-level API features and user-friendliness.

3.5. Model Training

This code is implementing a convolutional neural network (CNN) using the Keras API with TensorFlow as its backend. The architecture of the network is as follows:

The first layer is a convolutional layer with 16 filters (also called channels) and a kernel size of 3x3. The input shape is set to (28,28,3), which means that the input images are 28x28 pixels with 3 color channels (RGB). The activation function used is ReLU, and the padding is set to 'same', which means that the input image is padded with zeros so that the output size is the same as the input size.The second layer is another convolutional layer with 32 filters and a kernel size of 3x3. The activation function used is ReLU.The third layer is a max pooling layer with a pool size of 2x2, which means that the size of the output is half the size of the input. This helps to reduce the dimensionality of the feature maps and makes the network more computationally efficient.

The fourth layer is a convolutional layer with 32 filters and a kernel size of 3x3. The activation function used is ReLU, and the padding is set to 'same'.

The fifth layer is another convolutional layer with 64 filters and a kernel size of 3x3. The activation function used is ReLU.

The sixth layer is another max pooling layer with a pool size of 2x2 and padding set to 'same'.

The seventh layer is a flatten layer which converts the output of the previous layer into a 1D vector. This is necessary because the next layer is a fully connected layer which requires a 1D input.

The eighth layer is a fully connected layer with 64 neurons and an activation function of ReLU.

The ninth layer is another fully connected layer with 32 neurons and an activation function of ReLU.

Skin Cancer Classification Model

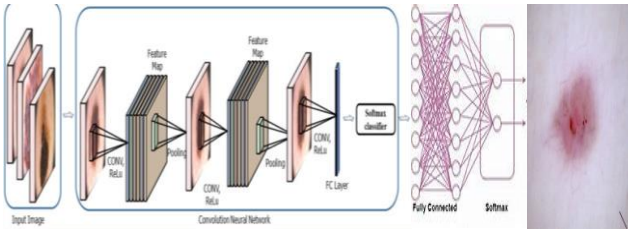

Figure 5

3.6. Implementation details

To better utilize both the first and the second moments of the gradient,this paper's model used the Adam optimizer. The Adam optimizer uses stochastic optimization and a constant learning rate.

Throughout the training process, the sparse categorical cross-entropy loss function was used as the loss function. This cross-entropy loss function us widely used loss function for classification. It coms from the binary cross-entropy, whre the lost L is calculated as in equation(1)

$$L= -\sum x\ [y\ln(p) + (1 – y)\ \ln(1 - p)] \quad (1)$$

In this formula, x is a sample point in the dataset, y is the label value, and p is the predicted probability of sample point x in class y. To expand this function to cover multi-class classification, the lost L is now calculated as in equation (2).

$$L = --\sum c -\sum\{x: y=c\}\ y\ln(p) \quad (2)$$

In this updated formula, c represents all the classes in the dataset, and p is the probability of the corresponding sample point in class c, while the meanings of the rest notations remain the same.
In addition, the training process set the epoch number to be 40 and the batch size of each epoch to be 256.

3.7. Model Evaluation

Model evaluation is an essential step in the development of any machine learning or statistical model. It helps to determine how well a model is performing on a given dataset, and whether it is able to generalize to new data.
The primary goal of model evaluation is to estimate the performance of a model, which can be done by comparing the predictions of the model to the actual values in the test dataset.

With each model, two graphs are obtained from the results of the training procedure. The Accuracy graph, which compares the training and validation accuracy plots, was one of them. The other was the Loss graph, which showed the training loss plot vs. the model's validation loss plot.

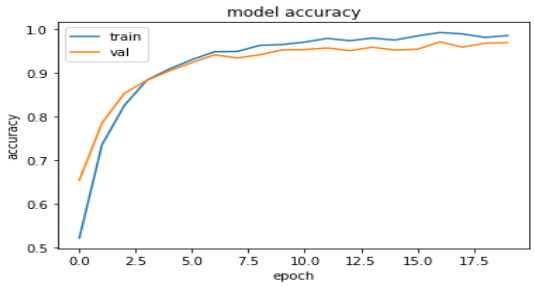

Figure 6: Model Accuracy

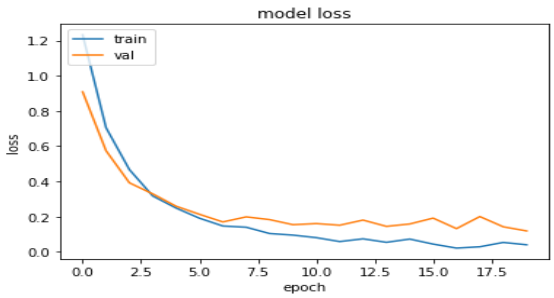

Figure 6: Model Loss

4. CHATBOT

The chatbot named I FOR YOU, provides a basic chatting function and can respond to the diagnosis based on the image uploaded by the user. Chatbots can be used in skin cancer classification by providing a conversational interface for users to input information about their skin condition and receive feedback on whether they should seek medical attention or not. The chatbot can use a combination of pre-defined responses and machine learning algorithms to classify skin lesions and provide information about their severity.One potential use case for a skin cancer classification chatbot is to help individuals perform self-assessments of their skin lesions. The chatbot can ask users questions about the size, shape, color, and other characteristics of the lesion, and use machine learning algorithms to classify it as benign or malignant. Based on the classification, the chatbot can provide recommendations on whether the user should seek medical attention or not.
Another use case for a skin cancer classification chatbot is to assist healthcare professionals in diagnosing skin lesions. The chatbot can be trained on a large dataset of skin lesion images and use machine learning algorithms to classify the images based on their features. The chatbot can then provide a recommendation on the diagnosis or treatment plan, which the healthcare professional can use as a reference.Overall, chatbots have the potential to be useful tools in skin cancer classification by providing a conversational interface for users to input information about their skin condition and receive feedback on the severity of the lesion.

4.1.Natural Language Processing(NLP)

Natural language processing is a field of computer science that deals with the development of computer systems that can be used to interpret text and speeches. If NLP is used to analyze words and to deal with text then certain aspects of language analysis and understand the use of machine learning. When machine learning is used for natural language processing and text analytics, it involves the machine-learning algorithm and artificial intelligence to understand the meaning of the documents and the text written.
As per the current needs, it seems that machine learning is an essential tool for natural language processing associated with the text classification and word sense analysis.[9]

4.2. Natural Language Toolkit (NLTK)

The Natural Language Toolkit (NLTK) is a widely-used open-source library for natural language processing in Python. It provides a comprehensive suite of tools and methods for working with human language data, including corpus management, text preprocessing, part-of-speech tagging, text classification, and linguistic analysis. With NLTK, researchers and developers can easily perform a range of NLP tasks, such as sentiment analysis, named entity recognition, and co-reference resolution. NLTK's modular design and ease-of-use make it accessible to both beginners and experts, and its active community provides support and a wealth of resources for users. NLTK has become a standard tool in the NLP field and is widely used in academia and industry

for a variety of applications, including chatbots, information retrieval, and text analysis.

Here, the chatbot first uses some methods to clean up the data. In the beginning part, we use the "sent_tokenize" and "word_tokenize" functions from the NLTK package to tokenize the words and sentences. After that, we clean up the tokens by Lemmatizing words by their root form, removing the stopwords the most frequent words used in Natural Language that are not useful for the text classifier, and removing the punctuations.

Then we define the function called "LemNormalize" to use the "LemTokens" function and the "remove punct-dict" dictionary that we defined for lemmatizing to clean and tokenize the text.After all these preparations we can proceed to the actual processing.We write a function called "start_chatbot" to packet all the chatting functions. It can call "greeting" to reply with greeting words politely including "hi", "hey", "*nods*", "hi there", "hello", "I am glad! You are talki ng to me","You're welcome, this is my job","You'd better talk with the doctor and you need further treatment".

4.3. TF-IDF

TF-IDF or Term Frequency–Inverse Document Frequency, may be a numerical statistic that's intended to reflect how important a word is to a document. Although it's another frequency-based method. It has two parts:

1.TF

TF stands for Term Frequency. It will be understood as a normalized frequency score. it's calculated via the subsequent formula:

TF=Frequency of word in a document / Total number of words

So one can imagine that this number will always stay ≤ 1, thus we now judge how frequent a word is in the context of all of the words in a document.

2.IDF

IDF stands for Inverse Document Frequency, but before we go into IDF, we must make sense of DF – Document Frequency. It's given by the following formula:

DF=Documents containing word W / Total number of documents

DF tells us about the proportion of documents that contain a certain word. So what's IDF? It's the reciprocal of the Document Frequency, and the final IDF score comes out of the following formula:

IDF = log(Total number of documents / Documents containing word W)

4.4. SpaCy

SpaCy is a relatively new Python NLP library designed for production use, making it more user-friendly than older libraries like NLTK. It boasts the fastest syntactic parser currently available on the market and is written in Python, making it efficient and quick. However, while it supports only seven languages compared to other libraries, the growing popularity of machine learning and NLP suggests that spaCy may begin to support more languages in the future.

spaCy is being used for text preprocessing, which involves tokenization, lemmatization, and removal of stop words and punctuations. spaCy is a powerful and efficient natural language processing (NLP) library that provides pre-trained models for a wide range of NLP tasks. The 'en_core_web_sm' model of spaCy, which is a small English model, is being used to preprocess the input text in this chatbot. The preprocess function in the code takes the input text, converts it to lowercase, tokenizes it, removes stop words and punctuations, and then lemmatizes each token. This results in a clean and normalized version of the input text, which can be used for further processing such as intent recognition, sentiment analysis, or information retrieval.

Using spaCy for text preprocessing has several advantages. First, spaCy is very fast and efficient, which is important for real-time chat applications. Second, spaCy provides accurate and reliable tokenization and lemmatization, which can improve the performance of downstream NLP tasks. Third, spaCy allows for easy customization and extension, which makes it suitable for building chatbots that can handle domain-specific language and jargon.

The chatbot needs to have access to a database of labeled examples and a trained classifier that can match the input text with the correct intent.Thus , we have created a file with intents named "symptom.txt" to collect the intents for the chatbot.

4.5. OpenCV"

cv2 is a Python library for computer vision tasks, specifically for working with images and videos. It stands for "OpenCV" (Open Source Computer Vision Library) version 2, which is a popular open-source computer vision and machine learning software library. It is used for various image processing tasks such as reading and writing images, image manipulation, color space conversion, filtering, feature detection, object detection and recognition, camera calibration, and many others. The library has a vast collection of functions and algorithms that are optimized to work on images and videos in real-time.

Here, a function called predict_skin_cancer that takes a model and a list of classes as input. The model is a trained neural network model for image classification, and the classes list contains the names of the classes that the model can predict.The function first asks the user to input the path of an image file. It then reads the image using OpenCV's cv2.imread function and displays the image using the cv2_imshow function from google.colab.patches module. The image is then resized to (28, 28) using OpenCV's cv2.resize function and converted to a floating point numpy array by dividing each pixel value by 255. The numpy array is then reshaped into the input shape expected by the model.The function then passes the

preprocessed image through the model to obtain a probability distribution over the classes. The index of the highest probability in the distribution is used to obtain the predicted class name from the classes list.
Finally, the function prints a message indicating that the picture is being processed and returns the predicted class name.

IV.CONCLUSIONS

This implements a chatbot named "I FOR YOU" to diagnose various skin cancer types. This helps the user to discover skin cancers at early stage and have a better chance of recovering. For the final result of the CNN model, the model receives an accuracy of rate around 97%. For NLP, the chatbot can handle basic conversations, including simple greetings, asking for input image, treatments and causes. In this era, chatbots paly a major role by intercting with the users in real time.

V. FUTURE WORK

A better data pre-processing may improve the result of the research. Thus, more pre- processing methods for images could be needed, including segmentation or boundary extraction. Also, the overfitting phenomenon is a problem that affects our results greatly. Merely applying the dropout method or adjusting the epoch size cannot solve this problem completely. Thus, more in-depth research is needed. And can connect the doctors or experts according to the location.Also, the chatbot is not intelligent enough, and one could bulid another neural network for a more intelligent NLP bot.

VI.Refrences

[1]. Diepgen, T. L., & Mahler, V. (2002). The epidemiology of skin cancer. British Journal of Dermatology, 146, 1-6.

[2] Jerant, A. F., Johnson, J. T., Sheridan, C. D., & Caffrey, T. J. (2000). Early detection and treatment of skin cancer. American family physician, 62(2), 357-368

[3] Mehta, Palak & Shah, Prof. (2016). Review on Techniques and Steps of Computer Aided Skin Cancer Diagnosis. Procedia Computer Science. 85.309-316. 10.1016/j.procs.2016.05.238

[4] Kassem, M.A.; Hosny, K.M.; Damaševičius, R.; Eltoukhy, M.M. Machine learning and deep learning methods for skin lesion classification and Diagnosis: A systematic review. *Diagnostics* 2021, *11*, 1390. [Google Scholar] [CrossRef]

[5] Bojarski, M., del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., ….& Zieba, K.(2016).End to end learining for self-driving cars.arXiv preprint arXiv:1604.07316

[6] Dildar, M., Akkkram, S., Irfan, M., Khan, H,U., Ramzan, M., Mahmood, A.R., Alsaiari, S.A., Saeed, A., Alraddadi, M., O., & Mahnashi, M.H.(2021). Skin cancer detection: A Review on Deep learning Techniques International journal of environmental research and public health, 18(10), 5479. https://doi.org/10.3390/ijerph18105479

[7] Karthik, R., Vaichole, T. S., Kulkarni, S. K., Yadav, O., & Khan, F. (2022). Eff2Net: An efficient channel attention-based convolutional neural network for skin disease classification. *Biomedical Signal Processing and Control*, *73*, 103406.

[8] Gujjar, P., & HR, P. K. (2022, December). Natural language processing using text augmentation for chatbot. In *2022 International Conference on Artificial Intelligence and Data Engineering (AIDE)* (pp. 248-251). IEEE.

[9] Sahana, M., & MM, S. G. NEWS CLASSIFICATION USING NATURAL LANGUAGE PROCESSING.

[10] Phayung Meesad1 , / Published online: 23 August 2021 © The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2021

1. [11] Nawaz, M.; Mehmood, Z.; Nazir, T.; Naqvi, R.A.; Rehman, A.; Iqbal, M.; Saba, T. Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering. Microsc. Res. Tech. 2022, 85, 339–351. [Google Scholar] [CrossRef] [PubMed]

[12] Abayomi-Alli, O.O.; Damasevicius, R.; Misra, S.; Maskeliunas, R.; Abayomi-Alli, A. Malignant skin melanoma detection using image augmentation by oversamplingin nonlinear lower-dimensional embedding manifold. *Turk. J. Electr. Eng. Comput. Sci.* 2021, *29*, 2600–2614. [Google Scholar] [CrossRef]

[13] Kadry, S.; Taniar, D.; Damaševičius, R.; Rajinikanth, V.; Lawal, I.A. Extraction of abnormal skin lesion from dermoscopy image using VGG-SegNet. In Proceedings of the 2021 Seventh International Conference on Bio Signals, Images, and Instrumentation (ICBSII), Chennai, India, 25 March 2021. [Google Scholar]

[14] Humayun, M.; Sujatha, R.; Almuayqil, S.N.; Jhanjhi, N.Z. A Transfer Learning Approach with a Convolutional Neural Network for the Classification of Lung Carcinoma. Healthcare 2022, 10, 1058. [Google Scholar] [CrossRef]

[15] Adegun, A.A.; Viriri, S. Deep learning-based system for automatic melanoma detection. IEEE Access 2019, 8, 7160–7172. [Google Scholar] [CrossRef]

[16] Afza, F., Sharif, M., Khan, M. A., Tariq, U., Yong, H. S., & Cha, J. (2022). Multiclass skin lesion classification using hybrid deep features selection and extreme learning machine. Sensors, 22(3), 799.

[17] Joy, A., Rani, S., & Sowmya, K. S. (2022, August). Classifying Skin lesions into Seven class using Deep Convolutional Neural Networks. In 2022 International Conference on Innovations in Science and Technology for Sustainable Development (ICISTSD) (pp. 264-268). IEEE.