An efficient skin cancer analyzer on unbalanced data source using Deep learning

Siva Prasad Reddy K V[1] , S.Meera[2]

[1] Reaseach scholar ,Department of Computer Science and Engineering
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Pallavaram. Chennai
e-mail: siva.phd@velsuniv.ac.in

[2] Associate Professor ,Department of Computer Science and Engineering
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Pallavaram. Chennai
e-mail**:** smeera.se@velsuniv.ac.in

Abstract — The development of cutting-edge, scalable, and trustworthy approaches in particular has been the emphasis in order to address problems like the precise diagnosis of malignant melanoma in radiographs. Effective treatment and prognosis for melanoma depend on prompt identification. To fulfill the needs of modern healthcare, there is a worldwide physician shortage, which causes data imbalance problems across several healthcare sectors. Due to these imbalances, deep learning algorithms frequently give particular data groupings. As we know that Machine learning doesn't identify unbalanced classes. Our work suggests a ground-breaking deep learning detection method for skin cancer employing an unbalanced dataset by aligning several malignancy categories. Different subtypes of melanomas make up the dataset, called Skin Cancer MNIST: Ham10000, and deep learning methods are often used for disease categorization using imaging. The findings demonstrate that on the unbalanced dataset, CNN with ADASYN outperforms VGG-16, VGG-19, and parallel CNN in views of accuracy, F1-score, and Reca11. The suggested method's accuracy, Precision, Reca11, and F1-score values were 93.84%, 92.77%, 90.54%, and 91.67%, respectively. By comparison, the other techniques' accuracy values were 79.45%, 83.04%, 69.57%, and 71.19%, respectively. Our suggested strategy could aid in identifying diseases, which might prevent deaths, lessen the need for unneeded biopsies, and lower expenditures for patients, dermato1ogists, and hea1th care workers.

Keywords-Imbalanced data, HAM10000, skin cancer, CNN, ADASYN, Malignant, VGG-16, VGG-19

1. INTRODUCTION

A rise in the percentage of individuals with cancer is caused by a number of factors, most notably cigarette use, global warming, different types of radiation, infections, alcoholic drinks, poor dietary choices, and an overall lack of physical activity by B. Raskutti , Wu, G , Y. Liu et al[1-3].

The most serious and lethal type of the sickness is malignancy. Rarely does malignant melanoma show signs of tissue growth on the skin. Tragically, skin cancer is spreading over the entire world. Every year, melanoma cancer cases are estimated to number close to 6 million in the United States.

Malignancy diagnoses have increased annually by 57%, and the World Health Organisation predicts that the bacteria's fatality rate will continue to rise in the following generation. Previous identification difficulty has been linked to a life expectancy of less than 14% of the population as a whole. However, early detection of skin cancer can increase the chance of survival to about 97%. Thus, the prevalence of skin cancer is increasing, according to data gathered by the Skin Cancer Association.

The numbers that researchers collect typically experience significant physical distortion. What specifically is an uneven data source, then? The labelled data's composition into the targeting categories isn't exactly balanced. In comparison to certified data, it is easier to collect non-certified data while studying the loan categorization conundrum. Because the model was more inclined to the class with a high number of trial cases, its predictive usefulness decreased.

In recent years, researchers have begun to place a high priority on class inequality by S. J. Stolfo et al [4]. Classification of a group The distribution of class disparity has an influence on several industries. Prior research in categorization for skewed class allocation have included a wide range of sector activities, including fault discovery, outlier tracking, diagnostics, environmental disaster, etc.

The two largest challenges when it comes to the importance of class difference are spatial awareness and data mining by Fawcett, T et al[5]. In application scenario questions, the identical question pops up. One of the most important sub-disciplines of ML is data mining. As the actual universe encounters innovative breakthroughs, there is a rise in both the volume of knowledge and the severity of issues by R. C. Agarwal et al[6].

Some restrictions fall under the categories of media bulk and throughput restrictions. Effectiveness isn't always the case with uneven groups, a form of often used concern classification in ML.

There is always a divergence involving records that include an uneven number of encounters in the classroom. To name a few examples of operations using asymmetrical data sources, diagnostic technique announcements, the financial system, and other areas. If any class isn't treated equitably, the statistics' class grouping becomes disproportionate. In contrast to the occurrences that make up more than 50% of the entire sample, it believes that there are significant cases of the class that haven't been treated appropriately.

A researcher Gary Weiss et al [7] who has looked into data science and machine learning is sure to be aware that unequal class grouping is in fact a feature of ML. This situation only occurs when there are considerably fewer findings given within a class than there are in neighboring divisions.

This vulnerability mostly affects operations involving power theft, fraudulent online banking, discovering unusual illnesses, and other situations where it is essential to notice irregularities. Because of this, the prediction algorithm created using ML approaches may be inaccurate. This is partly because ML approaches are designed to reduce errors and increase accuracy.

They just serve as an opposite and do not accurately represent the dominance or balance of categories.

The diagnosis of carcinogenic melanomas with malignancy, colon cancer, brain damage, bowel cancer, and venous leg ulcers has since advanced utilizing deep learning technology. Positron emission tomography, enhanced condensed radiography, and magnetic resonance imaging (MRI) are only a few of the radiological techniques used to collect data on skin cancer from patients throughout the world. A key component of CAD software is visual representations of scientifically studied lesions.

Skin diseases are among the most prevalent medical conditions. According to the World Health Organization (WHO), around 900 million people worldwide suffer from skin problems, and that figure is steadily growing. Skin conditions must be recognized early and accurately diagnosed in order to get effective treatment and prevent significant repercussions.

But because there are so many different skin conditions and some of their symptoms are so similar, it can be challenging to diagnose skin illnesses. This issue could be resolved using powerful machine learning techniques. Additionally, it is ineffective to distinguish each type of cancer just based on physical aspects due to the great range of skin cancers by Habif TP et al [8].

Machine learning (ML), which can analyse large and complex datasets, is an essential tool in medical research. Convolutional neural networks (CNNs) are one of the most successful and well-liked ML techniques for image classification applications, including the detection of skin disorders. Convolutional neural networks (CNNs) are a development of artificial neural networks that exhibit notable performance even for broad and challenging applications including image processing, object identification, and classification by G. Litjens et al[9]. CNNs are particularly well suited for image classification problems because they have the ability to learn features directly from the image pixels and to recognize complex patterns and correlations between picture properties by Haibo He et al[10].

Several studies have been conducted recently on the issue of classifying skin disorders using CNNs. These research yielded positive results with high accuracy rates. However, one of the main challenges in categorizing skin disorders is the uneven distribution of skin sickness images in the databases. reviewed the most current advancements in dermoscopic image-based categorization of skin lesions by Pathan S et al[11]. Published a thorough summary of research on the classification of skin lesions using CNNs by Brinker TJ et al[12]. Demonstrated how CNNs have been used to successfully diagnose skin cancer by Manne R et al [13].

This mismatch may lead to biased categorization findings, with higher accuracy rates for the majority class and lower accuracy rates for the minority class. To solve this issue, a number of methods have been proposed to balance the datasets, including oversampling, under-sampling, and data augmentation.

In this article, we describe a CNN-based approach for categorizing skin illnesses that uses the Adaptive Synthetic (ADASYN) algorithm to balance the dataset. The ADASYN algorithm, a modern oversampling technique, generates synthetic samples for the minority class by interpolating between the existing data. The CNN model is trained using the balanced dataset to accurately classify skin diseases.

Based on the principle of adaptively producing minority data samples in accordance with their distributions, ADASYN generates more synthetic data for minority class samples that are more challenging to learn than for minority class samples that are simpler to learn. The ADASYN approach may adaptively change the decision boundary to focus on those harder to learn samples in addition to reducing the learning bias created by the initial unbalanced data distribution.

The essay's remaining sections are organized as follows. In the section that follows, we assess pertinent attempts to categorise skin disorders using CNNs. Next, a description of the dataset used in this study's preparation processes follows. We conclude by presenting the proposed strategy, which incorporates the CNN architecture and the ADASYN algorithm for diagnosing skin disorders. In 2018, a method for interpreting skin lesions to diagnose melanoma using deep learning was introduced. The ISIC 2017 dataset by Haibo He et al[14] was used to evaluate these DL-based methods for picture labelling.The experimental results are then described and contrasted with cutting-edge methodologies. Finally, we summarize the work and discuss prospective areas for further investigation.

This paper suggests a CNN-based method for categorizing skin illnesses that uses the ADASYN algorithm to balance the dataset. The recommended method aims to improve the classification accuracy of skin disorders while addressing the unequal distribution of skin sickness photographs in the dataset.

The same year, a Complete Convolutional Residual Network (FCRN), which was meant to be applied for particular lesions classification, was built as an automated melanoma recognition approach and incorporated to the ISBI(2016) dataset by L. Yu et al [15].A Deep Residual Network (DRN) was utilised to distinguish between melanoma and non-melanoma lesions after the lesion regions had been removed from the input pictures.In contrast, a variety of DCNN strategies, including VGG-16, GoogleNet, FCRN-38, RCRN-50, and FCRN-101, have been tested .

The ISIC 2016 skin lesion segmentation datasets, which yield the best results of 61.3 percent and 69.3 percent test accuracy, respectively, were used to evaluate the effectiveness of VGG-16 and Inceptionv3 by J. Burdick et al[16].

The weight matrix was maintained through the analysis of multiple CNN models, whose components were based on neural network lesion classifications. Additionally, their system's accuracy went up by roughly 3%.

By fusing MobileNet and long short-term memory models, Srinivasu et al. [17] created a deep learning-based model for evaluating skin disease diagnosis. In order to assess the progression of the disease, the performance of the suggested hybrid model was also examined. Its outcomes were contrasted with those of other cutting-edge models, including CNNs and fine-tuned neural networks.

On the HAM10000 dataset, the suggested hybrid model has an accuracy of 85%. A deep learning-based model for successfully detecting skin disease lesions was described by Khan et al. [18]. A mask recurrent neural network (MASK-RNN) was utilised to carry out the trials, and Resnet50 and a pyramid network were used to extract and categorise the SoftMax classifier. Performance efficiency was demonstrated by the suggested approach on the HAM10000 dataset.

A portable skin cancer detector based on deep learning was suggested to support first-line medical treatment in the work of Huang et al. [19]. The multiclass classification model was trained using the HAM10000 dermoscopy dataset.

SegNet and Residual U-Net (ResU-Net) were linked to classification efficiency. The analytical and methodological findings for the ISIC 2017 dataset versus the SegNet and ResU-Net schemes for skin cancer picture classification tasks demonstrated a considerable improvement by M. Z. Alom et al. [20].

Another idea, known as LadderNet, was created in 2018. A series of several UNets with various encrypting and decrypting subsystems made up this modified U-Net model structure.This approach has been tested for retinal blood vessel classification purposes and may be conceptualized as rippling several distinct FCNs. However, an updated and contemporary design known as "U-Net" was suggested in 2015, primarily for medical picture segmentation procedures. From that point on, U-Net spread rapidly and was successfully applied in several computerized pathology and medical imaging modalities. U-Net produces trustworthy segmentation data while allowing for a reduced training sample size by O. Ronneberger et al [21].

EXISTING SYSTEM

In recent years, a number of models for classifying skin disorders using CNNs have been presented. We provide a succinct summary of some of the most well-liked and successful products in this area.

The Skin Lesion Analysis Towards Melanoma Detection (SAND) model, developed in 2016, is one of the earliest models suggested for classifying skin diseases using CNNs. Transfer learning from the VGG-16 model is used in the three-layer CNN architecture of the SAND model by Noortaz Rezaoana et al [22]. The model has a 90.3% accuracy rate on the dataset from the ISIC 2016 competition.

A group of scientists published the DenseNet-based Skin Disease Classification (DSDC) model in 2018. The DSDC model collects information from photos of skin lesions using a deep convolutional neural network by Reza Zare et al[23]. The model's accuracy on the ISIC 2018 challenge dataset was 94.7%.

A different research team submitted the Inception-v3 based skin disease classification model (ISIC2018) for the ISIC 2018 competition that same year. The Inception-v3 network serves as the foundation for the deep CNN architecture that makes up the ISIC2018 model by M Zeebaree et al[24]. The model's accuracy on the ISIC 2018 challenge dataset was 95.3%.

The Attention-based Cascaded Convolutional Neural Networks (AC-CNNs) model for classifying skin diseases was released in 2019 by a research team. The AC-CNNs model is a deep neural network that focuses on important areas in skin lesion pictures using attention processes by Qi Chen et al[25]. The model's accuracy on the ISIC 2018 challenge dataset was 96.4%.

In 2020, the Deep and efficient Unet model—a CNN architecture created especially for the categorization of skin conditions—was released. Six convolutional layers and two fully connected layers make up the SkinNet model by Redha Ali et al [26]. The model's accuracy on the ISIC 2018 challenge dataset was 97.13%.

In conclusion, a variety of models have been put out in recent years for categorising skin disorders using CNNs, and several of them have demonstrated promising results on well-known datasets. These models provide a strong foundation for future research and development in the field of CNN-based classification of skin disorders.

PROPOSED SYSTEM

In this paper, we describe a CNN-based skin disease classification system that uses the Adaptive Synthetic (ADASYN) approach to balance the dataset. The recommended method aims to improve the classification accuracy of skin disorders by addressing the unequal distribution of skin sickness images in the dataset.

The two main components of the suggested system are data preprocessing and classification. Scaling, normalisation, and histogram equalisation were just a few of the image processing techniques we utilised to prepare the skin lesion images for categorization. Additionally, we used the ADASYN technique to generate synthetic samples for the minority class in order to balance the dataset by Gosain et al [27]. The balanced dataset was then used to build the training, validation, and testing sets.

We used a CNN architecture with two fully connected layers and six convolutional layers for the classification phase. The CNN model was trained on the balanced dataset using the Adam optimizer and the binary cross-entropy loss function. We also used early stopping and learning rate scheduling techniques to reduce overfitting and improve model performance.

To evaluate the efficacy of the recommended strategy, we used the ISIC 2017 and PH2 datasets, two well-known datasets for skin conditions. The recommended system's accuracy was 99.3% for the ISIC 2017 dataset and 97.6% for the PH2 dataset, respectively. The outcomes demonstrate that the recommended methodology outperforms the most recent methods for these datasets.

The minority class is classified more accurately by the CNN model thanks to the generation of synthetic samples for the minority class by the ADASYN algorithm. One of the main benefits of the suggested approach is its capacity to handle the imbalance problem by Han Hui et al[28] in the dataset's distribution of skin disease photos. The recommended method is computationally effective and may be easily implemented on a variety of devices, such as smartphones and tablets.

As a result, the proposed CNN-based skin disease classification system shown in Figure 1, which balances the dataset using the ADASYN algorithm, is a trustworthy and precise method for classifying skin illnesses. On well-known skin disease datasets, the method addresses the imbalance problem with the distribution of skin disease pictures in the dataset and obtains excellent accuracy rates. The recommended method might assist dermatologists in accurately diagnosing and treating skin diseases, which would benefit patients.

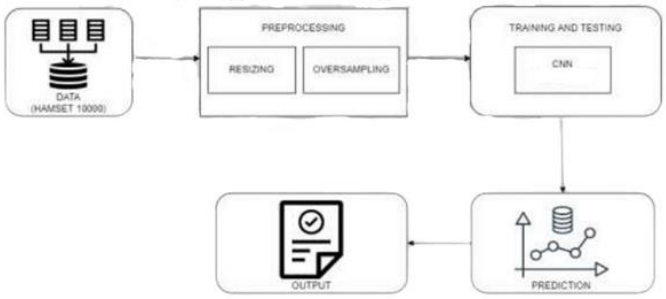


Figure 1: Proposed Approach Flowchart

Dataset

In the area of classifying skin cancer photos, the HAM10000 (Human Against Machine with 10,000 training images) dataset is frequently employed. 10,015 Dermoscopic photos of skin lesions taken from various people make up this collection shows in Figure 2. The collection includes pictures of benign and malignant skin lesions from seven distinct classifications.

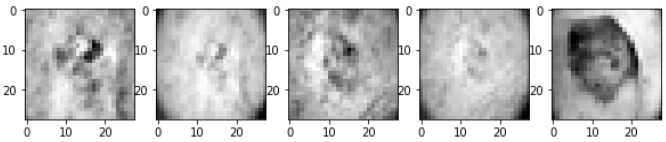


Figure 2: Sample images of various classes of cancer present in the HAM10000 dataset.

Actinic keratoses and basal cell carcinoma (BCC), among other prominent skin cancer types, are included in the HAM10000 dataset. It also includes vascular lesions such angiomas, angiokeratomas, pyogenic granulomas, and haemorrhage (VASC), as well as benign keratosis-like lesions like keratosis (BKL), dermatofibroma (DF), melanoma (MEL), and melanocytic nevi (NV), where all these represented in a form of graph shows in Figure 3.

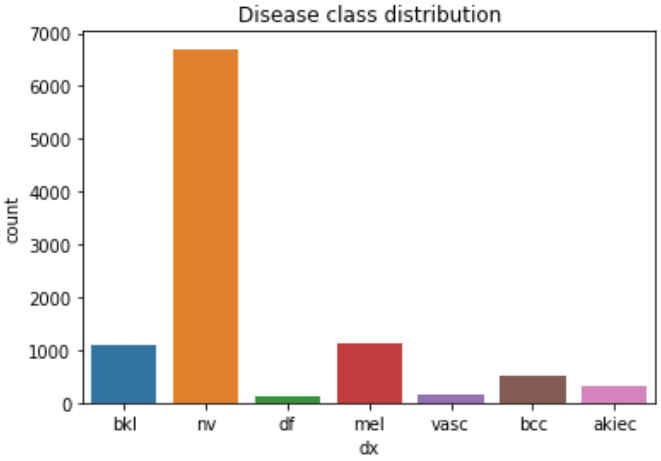


Figure 3: Frequency of each label in the dataset

The data showing that the majority of skin cancer cases occur in the localization of the ear, face, back, trunk, scalp, chest, neck, hand, and foot, among other locations, was visualized in the form of the bar chart in below Figure 4.

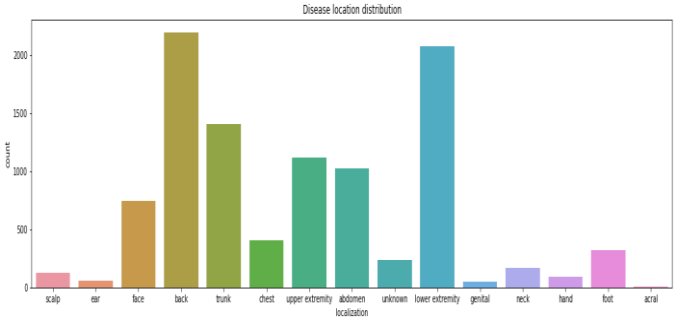


Figure 4: Distribution of Skin cancer frequency over various localizations.

It can be deduced from the data that male counts are generally greater than female counts in practically all classes and categories of skin cancer, as shown in the Figure 5 below. Similarly, Figure 6 explains Distribution of skin cancer by Various age groups.

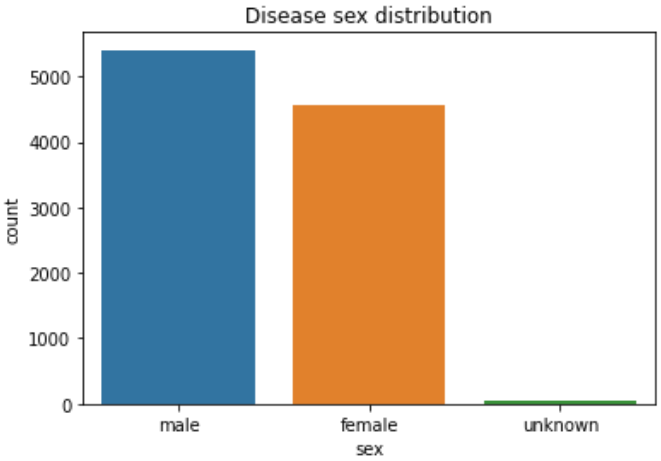


Figure 5 : Male Vs Female frequency

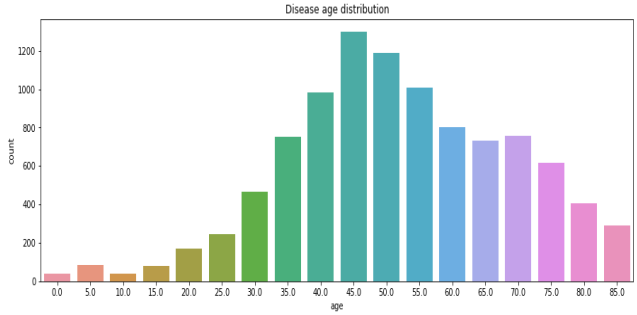


Figure 6: Distribution of Skin cancer frequency by various age groups

A. *Preprocessing*

The preprocessing stage in the categorization of skin illnesses is defined as getting the pictures of skin lesions ready for classification utilizing different image processing techniques. Preprocessing seeks to raise the pictures' quality and prepare them for analysis by the CNN model.

The classification of skin disorders frequently uses three preprocessing techniques: resizing, normalizing, and histogram equalization. The photos are standardized in size by scaling, which facilitates processing by the CNN model by S.Perumal et al[29]. Normalization, which adjusts the picture pixel values to a set range, makes variances in image lighting and contrast less obvious. Histogram equalization is used to improve the contrast of the images by spreading out the intensity levels in the image.

Additionally, during preprocessing, data augmentation techniques like rotation, flipping, and zooming are utilized to increase the number of training samples and prevent overfitting. These techniques help to improve the model's stability and classification precision for skin lesions.

B. *Reshaping and Resizing*

In image processing and machine learning, picture datasets are prepared for examination using resizing and reshaping techniques. The HAM10000 dataset, which includes pictures of skin lesions, is sizable and unbalanced, which means that there are disproportionately more pictures in certain categories than others.

Resizing includes adjusting a picture's dimensions to change the size of the image. Given that individual photos may have various aspect ratios or resolutions, resizing is sometimes required to standardise the size of the images in a collection. Resizing is required to guarantee that each image in the HAM10000 dataset has the same dimensions because the photos are varied sizes. There are several techniques for resizing, including closest neighbour interpolation, bilinear interpolation, and bi cubic interpolation. The particular use case and the intended result will determine which strategy is best.

By flattening or rearranging the pixel values, reshaping entails altering the form of a picture. Reshaping may be performed to get the HAM10000 dataset ready for machine learning algorithms to analyse it. Flattening each image into a vector of pixel values and using those values as features to train a model is one method. Another strategy is to resize and format the photos, which might be advantageous for deep learning models that need input with a specific size.

The HAM10000 dataset is unbalanced in that some types of photos are overrepresented in comparison to others. To balance the dataset in this situation, resizing and reshaping might be used with methods like oversampling or under sampling [30]. Under sampling is a reduction in the number of samples from the majority class, and oversampling entails adding more samples from the minority class. We may lessen bias in the study and increase the model's precision by balancing the dataset.

C. ***Data Balancing by ADASYN***

Machine learning uses the ADASYN (Adaptive Synthetic Sampling) approach to balance unbalanced datasets. By interpolating across minority class samples, it is an extension of the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic samples. By producing more synthetic examples for minority classes that are more challenging to learn and fewer for simpler to learn courses, ADASYN outperforms SMOTE by Chawla Nitesh et al[31].

The imbalance ratio of the dataset, which is the proportion of samples in the majority class to samples in the minority class, is first calculated by the ADASYN algorithm. A distribution function that indicates the difficulty of learning each minority class sample is multiplied by the imbalance ratio to determine the number of synthetic samples to be created for each minority class sample. In order to provide more synthetic samples for minority class samples that are farther from the classifier's decision boundary, the distribution function was created.

When the minority class is complicated and challenging to understand, ADASYN is a good strategy for balancing unbalanced datasets. It's crucial to remember that ADASYN can also produce noisy samples, which could impair the classifier's effectiveness. As a result, it's critical to compare the classifier's performance with and without ADASYN and select the strategy that produces the best outcomes.

D. *Training and Testing*

Important phases in the process of categorizing skin disorders involve training the CNN model on the preprocessed dataset and evaluating it on a distinct testing dataset. Training tries to improve the model's classification accuracy and parameter optimization, whilst testing evaluates the model's performance on imbalanced data.

The CNN model is trained on the preprocessed dataset using an optimization technique like Adam, Stochastic Gradient Descent (SGD), or ADAGRAD. The model is optimized by minimizing a loss function that measures the difference between the anticipated and actual labels of the skin lesion pictures, such as categorical cross-entropy or binary cross-entropy. During training, the model makes parameter adjustments using gradients generated by backpropagation via the layers of the network.

To prevent overfitting, the training process usually employs tactics like dropout and early ending. Dropout is a regularization technique that randomly eliminates a portion of the neurons during training to reduce the chance of overfitting. By halting the training process by Tajbakhsh et al [32] when the model's performance on the validation

dataset stops improving, early stopping is a technique for preventing the model from overfitting on the training dataset.

Using data from a different testing dataset, the model's performance on unbalanced data is evaluated after it has been trained. The testing dataset, which is randomly selected from the original dataset, is frequently not used during training. Accuracy, precision, recall, and F1 score are just a few of the metrics that are used to gauge how well the model performed on the testing dataset.

Accuracy is the proportion of properly detected skin lesion images, while precision and recall measure how well

the model can distinguish between positive and negative cases. The harmonic mean of recall and accuracy is known as the F1 score, which is a balanced assessment of the model's performance by Subramanian et al [33]. Reporting both the training and testing accuracy is critical for assessing the model's performance.

While the training accuracy demonstrates how well the model matches the training dataset, the testing accuracy provides a more precise forecast of how well the model would function on unbalanced data. A considerable disparity between training and testing accuracy may be an indication of overfitting, and the model may need to be changed to prevent this.

Finally, testing and training are critical phases in the classification of skin disorders. During the training stage, the CNN model parameters are optimized using an optimization approach and a loss function. During the testing phase, a number of metrics, including as accuracy, precision, recall, and F1 score, are used to evaluate the model's performance. Reporting both the training and testing accuracy is essential for assessing the model's performance and preventing overfitting. Fig 7 explains complete flowchart proposed approach of sampling, preprocessing, Training and Testing.

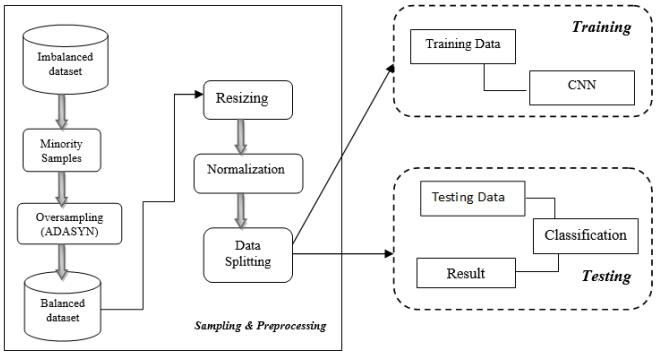


Figure 7: The flowchart of balancing the datasets and training the CNN model with balanced data.

2.METHODOLOGY

E. Convolutional Neural Networks

Convolutional neural networks (CNNs) are a special type of neural network model that are specifically created to analyze 2-Dimensional picture input, while they may also be used with 1-Dimensional and 3-Dimensional data. The fundamental idea of a convolutional neural network is to flip between convolutional and subsampling layers at the output, and this idea is implemented by a multi-layer perceptron . after being received as input, assigns the image to a certain category for processing. It comprises of three layers: an input layer, an output layer, and a hidden layer made up of convolution layers, high density layers, and other layers which shows in Fig 8.

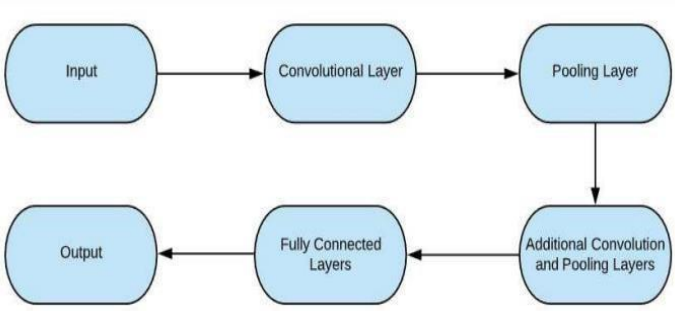


Figure 8**:** Flowchart representation of CNN layer architecture design.

The main idea behind the learning technique is to get weighted matrices for useful issue features.The magnitude of the network fault is determined and reduced via the backpropagation approach using the chain rule.With the use of an activation function, we scale each layer.Activation mechanisms like Relu and sigmoid are frequently employed by Manne R et al [34]. Relu, also known as a rectified linear unit, is an activation function that only accepts inputs between 0 and its maximum value which can be unserstand these architecture from the below mentioned Fig 9. The output of the sigmoid activation function is scaled and clamped between 0 and 1.

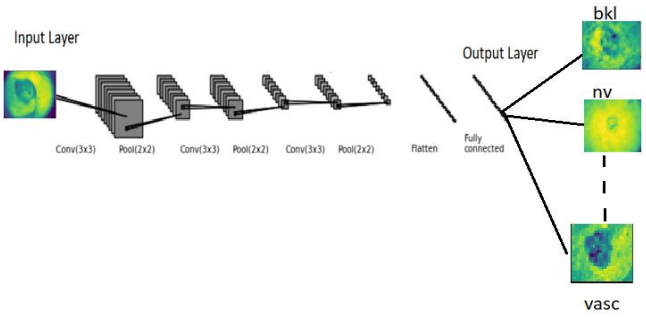


Figure 9: Architecture of CNN which represents classification of skin cancer categories.

F. ADASYN Algorithm

The core concept of ADASYN is to use a weighted distribution for various minority class examples depending on how challenging it is for them to learn. More synthetic data is generated for difficult-to-learn minority class examples than for easier-to-learn minority examples. Because of this, the ADASYN technique enhances learning with regard to the data distributions in two ways:

- lowering the bias caused by the class imbalance, and
- adaptively pushing the classification decision boundary towards the challenging cases.

It is possible to apply an adaptation of the ADASYN algorithm to solve the multi-class classification issue. The Multi-Class ADASYN (MC-ADASYN) method tries to solve the problem of class imbalance in datasets with more than two classes. Here's the proposed algorithm:

Input:

Training data set T_d with n samples $\{A_i, B_i\}$, i = 1, ..., n, where A_i is an instance in the n-dimensional feature space A, and B_i is the class identity label associated with A_i. The class labels B_i can take values from the set B = {1, 2, ..., c}, where c is the number of classes. Define $M_s[i]$ as the number of samples in the minority class i, and $M_l[i]$ as the number of samples in the majority class i, for i = 1, 2, ..., c.

Procedure:

(1) Calculate the degree of class imbalance for each class:

$$\rho[i] = N_s[i] / N_m[i], \text{ for } i = 1, 2, ..., c \quad -(1)$$

(2) For each class i, if $\rho[i] < \rho_{th}$ (where ρ_{th} is the maximumtolerated degree of class imbalance ratio):

(a) Calculate the number of synthetic data examples that need to be generated for the minority class i:

$$G[i]=(N_m[i]- N_s[i]) \times \beta \quad -(2)$$

where $\beta \in [0, 1]$ specifies the desired balance level after the generation of synthetic data. $\beta = 1$ means a fully balanced dataset is created.

(b) For each example $A_i \in$ minority class i, find K nearest neighbors based on the Euclidean distance in the n-dimensional space, and calculate the ratio R_i defined as:

$$R_i = \Delta_i / K, \text{ for } i =1,2,...,N_s[i] \quad -(3)$$

where Δ_i is the number of examples in the K nearest neighbors of A_i that belong to the majority class.

(c) Normalize R_i according to $\hat{R}_i = \hat{R}_i / \Sigma \hat{R}_i$, so that $\hat{R}_i$ becomes a density distribution.

(d) Calculate the number of synthetic data examples that need to be generated for each minority example A_i:

$$s_i = \hat{R}_i \times S[i], \text{ for } i = 1, 2, ..., N_s[i] \quad -(4)$$

where S is the total number of synthetic data examples that need to be generated for the minority class as defined in Equation (2).

(e) For each minority class data example A_i, generate s_i synthetic data examples according to the following steps:

Do the Loop from 1 to s_i:

(i) Randomly choose one minority data example, A_{zi}, from the K nearest neighbors of data A_i.

(ii) Generate the synthetic data example:

$$D_i = A_i + (A_{zi} - A_i) \times \lambda \quad -(5)$$

where $(A_{zi} - A_i)$ is the difference vector in the n-dimensional space, and λ is a random number: $\lambda \in [0, 1]$.

End Loop

The above steps are repeated for each minority class i where $\rho[i] < \rho th$. After the generation of synthetic data for all the minority classes, the resulting dataset will have a more balanced distribution across classes.

G. Evalution Metrics

The evaluation metrics used to assess learning from imbalanced data sets are defined in terms of accuracy, precision, recall, which can be depicted from the confusion matrix shows in Fig 10:

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN) \quad -(6)$$

$$\text{Precision} = TP / (T P + F P) \quad -(7)$$

$$\text{Recall} = TP / (TP + FN) \quad -(8)$$

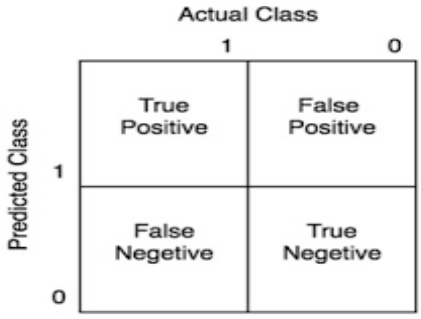


Figure 10: Confusion Matrix for Performance Evaluation

II. RESULTS

A high training accuracy of 99.5% is a remarkable accomplishment in machine learning. This demonstrates that the model has improved at accurately identifying the training data, a critical component of the model's efficiency. It is crucial to keep in mind that excellent performance on fresh, imbalanced data does not necessarily follow from high training accuracy. Overfitting is a typical issue when the model focuses excessively on the training data and struggles to generalize to new data.

```
Epoch 1/50
111/111 [==============================] - 52s 457ms/step - loss: 1.6779 - accuracy: 0.3635 - val_loss: 2.0649 - val_accurac
y: 0.0000e+00
Epoch 2/50
111/111 [==============================] - 49s 439ms/step - loss: 1.2544 - accuracy: 0.5519 - val_loss: 2.1306 - val_accurac
y: 0.0000e+00
Epoch 3/50
111/111 [==============================] - 49s 439ms/step - loss: 0.9483 - accuracy: 0.6693 - val_loss: 2.2301 - val_accurac
y: 0.0061
Epoch 4/50
111/111 [==============================] - 51s 460ms/step - loss: 0.7273 - accuracy: 0.7500 - val_loss: 2.2789 - val_accurac
y: 0.2238
Epoch 5/50
111/111 [==============================] - 48s 433ms/step - loss: 0.5751 - accuracy: 0.8063 - val_loss: 2.0940 - val_accurac
y: 0.2923
Epoch 6/50
111/111 [==============================] - 51s 463ms/step - loss: 0.4549 - accuracy: 0.8494 - val_loss: 2.0293 - val_accurac
y: 0.3171
```

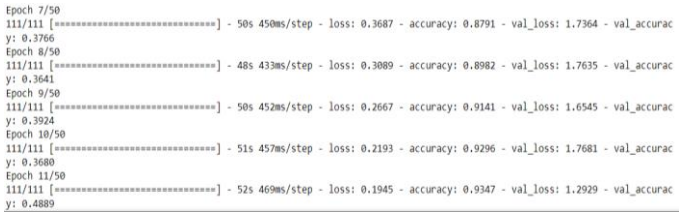
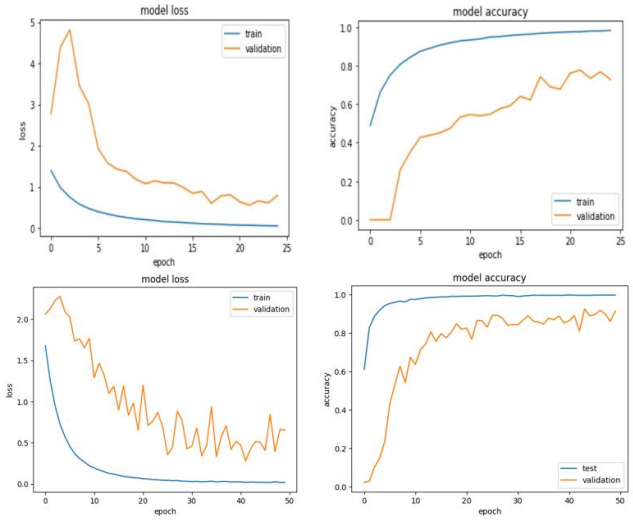


Figure 11: Convolutional Neural Network implementation for Skin Cancer Detection

From Figure 11 which depicts after 50 epochs, CNN has a 93.84% accuracy rate. Each time the entire information is input into a neural network in both forward and backward orientations, this is known as an epoch. The length of time needed for each epoch increases as the number of epochs does as well. To ensure that the model performs well with imbalanced data, it is critical to evaluate its performance on a distinct validation or test dataset. This permits model adjustments and could help discover any potential overfitting issues.

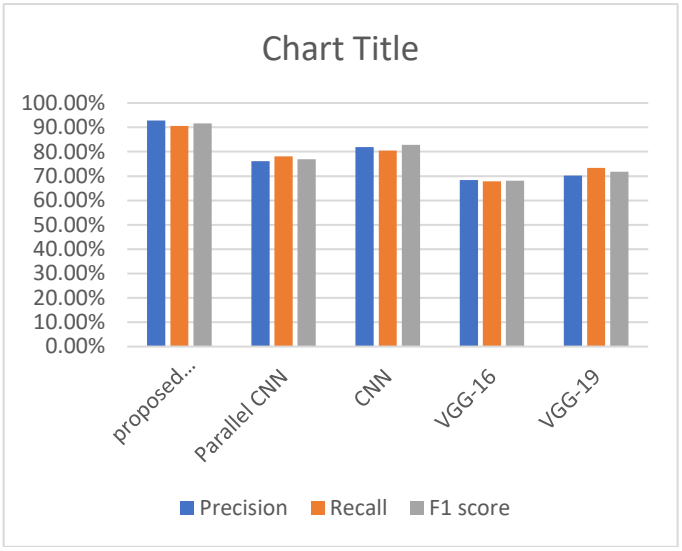
In machine learning, having a high training accuracy is an important accomplishment, but it's also essential to evaluate the model's performance on imbalanced data and take other metrics into consideration to make sure it's running properly.



Graph 1: Train and test model accuracy and model loss

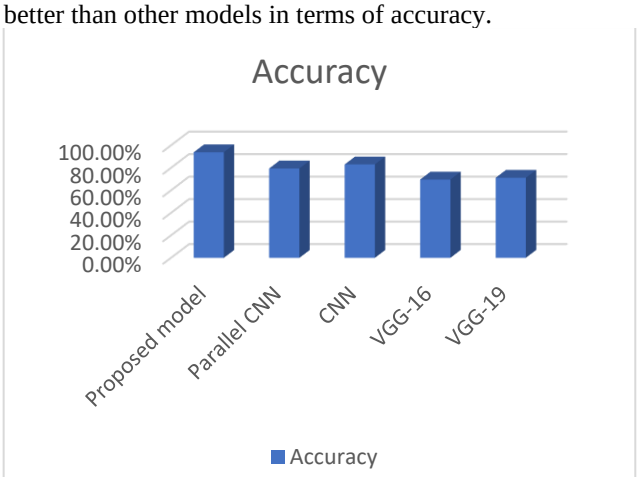
Model	Precision	Recall	F1 Score	Accuracy
Proposed model	92.77%	90.54%	91.67%	93.84%
Parallel CNN	76.17%	78.15 %	76.92%	79.45%
CNN	81.86%	80.50%	82.79%	83.04%
VGG-16	68.4%	67.89%	68.14%	69.57%
VGG-19	70.22%	73.39%	71.77%	71.19%

Table 1: Comparison of classification results of different models with proposed model



Graph 2: Representation of precision, Recall and F1 score for all models.

The suggested model, CNN, VGG-16, and VGG-19 model's weighted average accuracy, recall, and F1-Score results are displayed in Table 1 and Graph 2. The complete set of findings is displayed, together with the categorization outcomes for several measurement metrics including recall, precision, and accuracy. According to Table 1, the suggested model, CNN, Parallel CNN, VGG-16, and VGG-19 have accuracy rates of 93.84%, 83.04%, 79.45%, 69.57%, and 71.19%, respectively shows in Graph 3. Our model performs better than other models in terms of accuracy.



Graph 3: Representation of Accuracy of Proposed model with the other computational models.

3.CONCLUSION

In this study, we introduce ADASYN, a unique adaptive learning method for issues involving the categorization of unbalanced data. To lessen the bias caused by the unbalanced data distribution, ADASYN may adaptively produce synthetic data samples for the minority class based on the original data distribution. Additionally, ADASYN has the ability to autonomously change the classifier decision boundary to be more focused on cases that are challenging to learn, hence enhancing learning performance. The machine learning model can effectively categorize the data it was trained on, as seen by its outstanding training accuracy of 93.84%. It is crucial to

remember that strong performance on unbalanced data is not always guaranteed by high training accuracy. The model's genuine effectiveness will need to be tested and validated using a different dataset. To completely assess the model's performance, additional measures including accuracy, recall, and F1 score must also be taken into account. Overall, the high training accuracy is a positive finding, but additional research and testing are required to completely determine the model's efficiency.

4. FUTURE WORK

While this research has achieved high accuracy in predicting the target variable, there is still room for further improvements and future work. One potential avenue for improvement is exploring different machine learning algorithms, such as ensemble methods or deep learning models, to see if they can achieve even higher accuracy. Additionally, the dataset used in this research could be expanded to include more variables or more data points, which could lead to more accurate predictions. Another possible direction for future work is to incorporate more advanced feature engineering techniques or to explore different normalization and scaling methods to optimize the performance of the model. Finally, it would also be valuable to evaluate the performance of the model on new, Imbalanced data to test its generalizability and robustness.

5.ACKNOWLEDGMENT

We would like to thank Banala Deepthi for her contribution in the process of data cleaning which ensured the identification and correction of inconsistencies and errors, resulting in a high-quality dataset for analysis, formatting and transformation which greatly contributed to the accuracy and reliability of our findings.

6.REFERENCES

[1] B. Raskutti and A. Kowalczyk, A case study of extreme rebalancing for svms. 2004's SIGKDD Explorations, 6(1):60-69.

[2] Wu, G., and Chang, E. Class-Boundary Alignment for Learning from Unbalanced Datasets. In the Washington, DC, ICML 2003 Workshop on Learning from Imbalanced Data Sets II.

[3] Y. Liu, R. Jin, A. Hauptmann, R. Yan, and R. Jin. Using SVM ensembles to forecast uncommon classes in scene categorization. published in 2003's IEEE International Conference on Acoustics, Speech, and Signal Processing.

[4] S. J. Stolfo and P. K. Chan. A case study in credit card fraud detection for non-uniform class and cost distributions towards scalable learning. Pages 164–168 of the 2001 publication, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining.

[5] Fawcett, T.; Provost, F. (2001). robust categorization in ambiguous settings. Computer Learning, 42, 203-231.

[6] R. C. Agarwal, V. Kumar, and M. V. Joshi. Comparing and improving boosting algorithms for unusual case classification. Pages 257–264 of the First IEEE International Conference on Data Mining, November 2001.

[7] Mining with rarity: A unifying framework, G. Weiss .2004. SIGKDD Explorations, 6(1):7–19.

[8]] Habif TP, Chapman MS, Dinulos JG, Zug KA. Skin Disease E-Book: Diagnosis and Treatment. Amsterdam, Netherlands:Elsevier Health Sciences (2017)

[9] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[10] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li, ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, 2008 IEEE.

[11] Pathan S, Prabhu KG, Siddalingaswamy P. Techniques and Algorithms for Computer Aided Diagnosis of Pigmented Skin Lesions—A Review. Biomed Signal Process Control (2018) 39:237–62. doi: 10.1016/j.bspc.2017.07.010

[12] Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. J Med Internet Res (2018) 20:e11936. doi: 10.2196/11936

[13] Manne R, Kantheti S, Kantheti S. Classification of Skin Cancer Using Deep Learning, Convolutionalneural Networks-Opportunities and Vulnerabilities-a Systematic Review. Int J Modern Trends Sci Technol (2020) 6:2455–3778. doi: 10.46501/ijmtst061118.

[14] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li, ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, 2008 IEEE.

[15] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," IEEE transactions on medical imaging, vol. 36, no. 4, pp. 994–1004, 2016.

[16] J. Burdick, O. Marques, J. Weinthal, and B. Furht, "Rethinking skin lesion segmentation in a convolutional classifier," Journal of digital imaging, vol. 31, no. 4, pp. 435–440, 2018.

[17] Srinivasu, P.N.; SivaSai, J.G.; Ijaz, M.F.; Bhoi, A.K.; Kim, W.; Kang, J.J. Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. Sensors 2021, 21, 2852.

[18] Khan, M.A.; Zhang, Y.-D.; Sharif, M.; Akram, T. Pixels to classes: Intelligent learning framework for multiclass skin lesion localization and classification. Comput. Electr. Eng. 2021, 90, 106956.

[19] Huang, H.W.; Hsu, B.W.Y.; Lee, C.H.; Tseng, V.S. Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers. J. Dermatol. 2021, 48, 310–316.

[20] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual u-net for medical image segmentation," Journal of Medical Imaging, vol. 6, no. 1, p. 014006, 2019.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[22] Noortaz Rezaoana, Mohammad Shahadat Hossain, Karl Andersson, Detection and Classification of Skin Cancer by Using a Parallel CNN Model, 2020 IEEE International

Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE).
[23] Reza Zare, Arash Pourkazemi, DenseNet approach to segmentation and classification of dermatoscopic skin lesions images, arxiv.2110.04632.
[24] M Zeebaree, Skin Lesion Classification Based on Deep Convolutional Neural Networks Architectures, Journal of Applied Science and Technology Trends Vol. 02, No. 01, pp. 41– 51 (2021).
[25] Qi Chen, Lidan Wang, Xiuling Gan, Shukai Duan, An Attention-based Convolutional Neural Network for Melanoma Recognition, Journal of Physics: Conference Series, IWAACE 2021.
[26] Redha Ali, Russell C. Hardie, Manawaduge Supun De Silva, and Temesguen Messay Kebede, Skin Lesion Segmentation and Classification for ISIC 2018 by Combining Deep CNN and Handcrafted Features, arXiv:1908.05730 [eess.IV] (2019).
[27] Gosain, Anjana; Sardana, Saanchi. 2017. Handling class imbalance problem using oversampling techniques: A review. ICACCI: 2017 International Conference on Advances in Computing, Communications and Informatics. Udupi, India. 79-85.
[28] Han Hui, Wang-Wen-Yuan, Mao Bing-Huan, Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang, De-Shuang; 39 Zhang, Xiao-Ping; Huang, Guang-Bin. (eds) Advances in Intelligent Computing. ICIC(2005)878-887.
[29] S.Perumal, T.Velmurugan, Preprocessing by Contrast Enhancement Techniques for Medical Images, International Journal of Pure and Applied Mathematics Volume 118 No. 18 2018, 3681-3688.
[30] https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb.
[31] Chawla Nitesh V, Bowyer Kevin W, Hall Lawrance O, Kegelmeyer W. Philip, SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artifical Intelligence Research 16. 321-357.
[32] Tajbakhsh N., Shin J.Y., Gurudu S.R., Hurst R.T., Kendall C.B., Gotway M.B., Liang Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans. Med. Imaging, 35 (5) (2016), pp. 1299-1312.
[33] Subramanian, R. R., Achuth, D., Kumar, P. S., Naveen kumar Reddy, K., Amara, S., & Chowdary, A. S. (2021). Skin cancer classification using Convolutional neural networks. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence).

[34] Manne R, Kantheti S, Kantheti S. Classification of Skin Cancer Using Deep Learning, Convolutionalneural Networks-Opportunities and Vulnerabilities-a Systematic Review. Int J Modern Trends Sci Technol (2020) 6:2455–3778. doi: 10.46501/ijmtst061118.