

Comparative Analysis of Skin Cancer and Skin Lesion Classification: A Performance Evaluation of CNN Image Classifiers

Abu Bakar Hasnath, Faisal Shahriar, Sayad Md. Prio, Sulaiman Hossain Tonmoy,
Annajiat Alim Rasel

Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh
abu.bakar.hasnath@g.bracu.ac.bd, faisal.shahriar@g.bracu.ac.bd,
sayad.md.prio@g.bracu.ac.bd, sulaiman.hossain@g.bracu.ac.bd,
annajiat@gmail.com

Abstract—Abnormal growth of skin cells, commonly known as skin cancer, can form life-threatening diseases if not treated timely. This necessitates proper identification and immediate treatment of skin cancer. Recently, there has been a notable advancement in using machine learning methods to analyze medical images. There is a huge potential in machine learning methods for automatically detecting and categorizing skin cancer. This paper explores a comprehensive technique using machine learning tools for detecting and categorizing skin cancers. For this project, we have used the HAM10000 dataset which consists of 10000 medical images of skin cancer varying in types and patient demographics. Our project analyzes the efficiency of InceptionV3, ResNet50, VGG16. Rigorous assessment methods like Accuracy, precision, Recall rate, and F1-score are used to measure the effectiveness of these algorithms. Overall model performance was strong with Inceptionv3 acquiring the highest accuracy of 83.63%, while InceptionResNetv2 gave an accuracy of 82.70%. Our presented approach makes a significant contribution to advancing dermatology by creating advanced diagnostic tools that will be crucial for healthcare professionals and also beneficial for patients affected by skin cancer. Our investigation shows the potential of Machine learning methods to enrich health outcomes by enhancing automated early detection processes enabling us with personalized treatments tailored to effectively address skin cancer.

Index Terms—Skin cancer; Skin lesion; AI; InceptionResNetV2; Inceptionv3; ResNet50; VGG16; Vision Transformer; explainable ai;

I. INTRODUCTION

Skin cancer is a common disease which is caused by abnormal growth in skin cells, sometimes by extended exposure to UV radiation emitted from the sun or tanning beds. There are three prominent types of skin cancer, which are cell Carcinoma, Squamous cell carcinoma, and Melanoma. Basal cell carcinoma and squamous cell carcinoma mainly grow on the skin which is exposed on the sun like face, neck and hands. These types can be cured generally when it can be identified in an early stage. On the other hand melanoma which is very uncommon and aggressive, which can grow in other body regions unless it is not properly treated.

Nowadays doctors use machine learning algorithms that can help them to cure skin cancer by improving early detection.

These algorithm use extensive data analysis which includes the examination of moles and lesions in order to detect patterns which shows them the presence of cancerous development. Machine learning models can effectively identify dangerous lesions and help doctors to take effective diagnosis by training them on various kinds of skin images. In addition, these algorithms can select patients with a high risk of developing skin cancer for extra inspection includes advance intervention and potentially slowing the growth of the disease. The introduction of machine learning algorithm in medical field improves efficiency and accuracy therefore, patients can easily get cured from skin cancer.

Precise classification of skin lesions and cancers is essential in the field of dermatological to ensure accurate diagnosis and effective treatment. Moreover, having the outrageous expense of hardware, it is crucial to locate a machine learning (ML) model that is economical. This paper presents a thorough examination of the HUM10000 dataset which can evaluate the effectiveness of conventional machine learning models including Decision Tree, Random Forest, KNN (K-Nearest Neighbors), and SVM (Support Vector Machine). Furthermore, it assesses the effectiveness of four cutting-edge image classification models—InceptionResNetV2, Inceptionv3, VGG16, and ResNet50—in accurately categorizing skin conditions. This study aims to analyze both Convolutional Neural Networks (CNNs) and traditional transformer-based approaches in order to determine the most efficient and cost-effective methods for classifying skin cancer and skin lesions. [7] Comparative analyses are useful references for medical practitioners and researchers who are looking for the best solutions in medical diagnostics.

II. LITERATURE REVIEW

In this research paper [1], Hossian et al. combined three different convolutional neural networks techniques. In this way, the authors of the paper were able to identify a normal kidney image or include a cyst, stone or tumor. The authors divided the images into sections according to the impacted region with the help of a watershed algorithm. Models like EANet,

ResNet50, and a modified convolutional neural network were employed. The two other versions, EANet and ResNet50, achieved 83.65% and 87.92% accuracy, respectively, while the modified version achieved the highest accuracy of 98.66%.

The Pyramid GNN model was presented by the authors in [2]. as a way to categorize COVID-19 cases from chest X-rays. In the paper, the author's suggested model uses Convolutional Neural Network(CNN) in splitting chest X-ray images into patches. After that, those patches are processed to produce a feature vector. These patch features are then interpreted as nodes in a graph structure in pyramid GNN. Next the authors applied graph convolution procedures where the nodes in the GNN model exchange information with neighboring nodes which improves the features even further. Finally, the authors used a multilayer perceptron which uses the features from the earlier GNN layers to classify COVID-19 cases. In this paper, the authors used three distinct CXR picture datasets. The authors lastly compared their suggested model to other models and assessed their suggested model showed better accuracy.

SVM was utilized by Noor et al. [3] for the classification of skin cancer in cases where the dataset is unknown. Several machine learning techniques, including convolutional neural networks (CNNs), artificial neural networks (ANNs), inception V3s, and many more, have been employed for the classification of malignancies and the diagnosis of melanoma. To characterize images, support vector machines (SVMs) have been employed. Using an ANN method, this research article achieved 96% sensitivity and 93% specificity.

The architecture of convolutional neural networks (CNNs) for early lung cancer detection and diagnosis is covered by Tiwari et al. [4]. To evaluate harmful knobs, the authors proposed a new paradigm based on profound learning. They employed faster R-CNN for nodule recognition and CMixNet for lung nodule detection. For nodule characterisation, the gradient boosting machine (GMB) is applied to the outlines of the planned 3D CMixNet arrangement. In order to lower false positives, the authors in this paper combined pathogenesis with clinical symptoms. In this paper the authors used a dataset of CT scan images of lung cancer patients. They preprocessed the dataset by applying thresholding, downsampling, and standardization techniques. Additionally, the authors used watershed segmentation to increase segmentation accuracy to separate lung tissue and nodules.

In this research paper Sea-Lim et al. [5] presented a unique approach to classify skin lesions. Here the authors used a lightweight Convolutional Neural Network (CNN) called MobileNet. Furthermore, the authors utilized two iterations of mobileNet one original and the other one modified version to investigate the dermatoscopic image data from the HAM 10000 dataset.

A deep neural network was trained to categorize skin lesions in a study by Esteva et al., and the results were compared to dermatologists' performance [6]. According to the authors, the effectiveness of the neural network was assessed through the use of sensitivity and specificity metrics. The area under the curve (AUC) metric was employed to compare the

network's findings with those of the dermatologists. For both tasks, the neural network's AUC was greater than 91%. The authors suggested that deep neural networks may be able to classify skin cancer with an accuracy level that is on par with dermatologists.

III. DATASET

A. Data Collection

The data used for this comparison is a set of pictures of pigmented skin spots with high quality taken from the HAM10000 set. The HAM10000 set has 10,000 pictures showing different types and sizes of spots as well as details about the patients. This makes it great for teaching and testing machines to classify skin problems. [7]

B. Data Visualization

a) Age Distribution: The assessment of age distribution in the HAM10000 dataset has been conducted with the aim of comprehending the demographic features of the specimen. Examination results unveiled a notable concentration of specimens within the age bracket of 40 to 50 years, hinting at the dataset's inclusion of a broad spectrum of ages.

b) Disease Occurrence: A thorough examination was undertaken to ascertain the prevalence of each category of skin lesions in the dataset. The inspection unveiled a notable disparity among the classes, comprising 6405 instances of melanocytic Nevi and a mere 110 cases of dermatofibroma. This revelation underscores the necessity to rectify the class imbalance issue in both the training and assessment phases of the model development process.

c) Gender Representation: A bar chart was used to visualize the distribution of gender within our dataset. Figure 1c shows the gender distribution of the individuals whose skin lesions were imaged.

d) Correlation between Age and Skin Lesions: A scatter plot was used to indicate the correlation between age and different skin lesions visually. Figure 2 demonstrates these correlations.

e) Localization of Lesions: An assessment of the allocation of lesion placements was performed to ascertain the prevailing sites attributed to skin lesion manifestation. This study gives important information about the body areas where skin problems are most likely to happen. It helps us understand how these problems spread on the body.

C. Data Augmentation

Data Augmentation was applied to make the dataset more robust and balance the class imbalance problem. Techniques for data augmentation were used for:

a) Image Augmentation: We used rotation, flipping, zoom, and shift like image augmentation techniques to induce multiple variants of the dataset apart from compensating the class imbalance issue.

b) Synthetic Data Generation: To fix the uneven data, we made new examples for the smaller group. We used SMOTE (Synthetic Minority Over-sampling Technique) to create more samples for the group with fewer examples. This method makes new samples from the group that is not represented well.

IV. METHODOLOGY

A. Data Preprocessing

Data preprocessing consists of a variety of workflows to bring the data to a position where it can be utilized to test the models. Data needs to be cleaned up before it heads to utilization and balance as if it is an imbalanced dataset. The dataset has a lot of irrelevant parts like shadow and sizing issues as each one may not be the same size but while giving it to test the model, we need to make sure all the data is of the same size. The dataset was cleaned manually, and the Tomato disease dataset contained some corrupt and irrelevant images. OpenCV is used for resizing the image to reduce computational load and ensure consistency. The target size of the resizing image was 224×224 . Thereafter, normalizing the image pixels into a common one so that every image is a standard size which will allow testing the data in a structured manner. Moreover, rotating the data in the range of 30, width shift range 0.2, shear range 0.2, zoom range 0.2 and flip mode was set to nearest so that every pixel is of the same rotation and zoomed and shifted in a right direction, which will ensure augmentation of the data. These parameters were set using the ImageDataGenerator class imported from the Keras library. [?] Soon after that oversampling or undersampling was utilized to target the imbalance data into a balanced one. After that, proportioning the data for training to 80% of total images, validation 10%, and testing 10% to see the result according to it and to compare how the models are working. Lastly, The quality of the data is worth working with or not and needs to be checked.

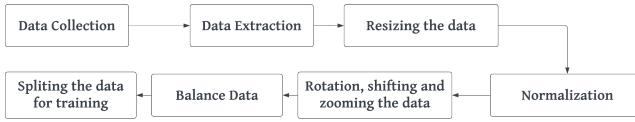


Fig. 1. Preparation phase of Data

B. Models:

a) InceptionResNetV2: InceptionResNet V2 is a deep convolutional neural network architecture which combines two architectures Inception and ResNet architecture. It is developed by Google's research architecture team. Inception modules are specialized on multiple convolutional networks of different sizes with the same layer. It works best in utilizing the resource to capture both local and global features. A key feature, Residual connection involves adding skip connections that allow gradients to propagate more effectively during training. This architecture stacks multiple inception block, which

contains parallel branches of convolutional with different kernel size ($1*1$, $(3*3)$, and $(5*5)$). With each inception block, residual connection intercorporate. This network typically ends with global pooling and connected layers and a softmax layer for classification. However, during its development, Inception-ResNet v2 demonstrated state of art performance on various benchmark datasets for image classification and computer vision task implementation. [?]

b) InceptionV3: This architecture is the evolution from InceptionResNet V2. This module allows the network to capture features at different scales simultaneously. By using filter sizes in parallel, the network can capture features at different scales and complexity allowing for richer presentation. However, inception modules are the key component of Inception architecture. Moreover, Auxiliary classifiers are also added to mitigate the layers during the training process to reduce the vanishing gradient problem. This module begins with standard convolution layers for initial feature extraction. This architecture is widely used for classification, object recognition, and feature extraction.

c) ResNet50: ResNet50 is a deep convolutional neural network architecture, which works in depth and residual connections. It is developed by the Microsoft Research team. It is a part of ResNet and contains 50 years. Here ResNet50 contains residual blocks and each block contains several layers. Residual block consists of skip connections that can bypass one or more layers. In the block, a bottleneck architecture is used, which is $(1*1)$, $(3*3)$, $(1*1)$. The $(1*1)$ reduce the dimensions, $(3*3)$ convolutions capture spatial information, and $(1*1)$ expand the dimensions. This architecture's design enables the training of deeper neural networks. ResNet50's balance of depth and computing efficiency makes it, along with other ResNet versions, a popular choice for image classification tasks, transfer learning, object identification, and image feature extraction. In the field of deep learning, its design has become a standard for a variety of vision-related applications. [?]

d) VGG16: VGG16 is convolutional neural network architecture. It is developed by Visual Graphic Group (VGG) at University of Oxford. It consists of 16 layers and 13 convolutional layers and 3 fully connected layers. This architecture, have multiple convolutional layers with $3*3$ filter, which is situated one after one. Moreover, max pooling layers with $2*2$ filters and 2 are convolutional layers to reduce spatial dimensions and dominant features. These layers are followed by fully connected layers at the end of the network. Firstly, VGG16 is initiated with an input layer, where the image is taken inside. The network is constructed from up of numerous convolutional layer stacks. Every stack has a max pooling layer after it. As data passes through the network, convolutional layers perform the function of feature extractors, extracting low-level to high-level features from the input image. Max pooling layers preserve the most significant features in the feature maps while decreasing their spatial dimensions. The feature maps are flattened and sent through fully connected layers at the conclusion of the network, where they are classi-

fied using the characteristics that have been learnt. However, VGG has a simple architecture with small convolutional layers, which makes it easy to understand. [?]

Each of the models is run through the HAM10000 dataset and compared with each one to have a clear view upon the detection process which will be best fit to use it. Enhancing the data and splitting those into train, validation and test is quite a necessary step to utilize the models. After that, separating the generator and validating it for text data is one of the essential steps for model validation. Soon after that, a base model is implemented according to which domain it is working on, however customizing it adds new values to it. When testing ends, the learning phase of the model starts to grab the changes according to the model run. Moreover, tuning is one of the essential steps of the model processing to have specified results. Furthermore, evaluation of the model is needed to confirm how this model is working and how all the other aspects of the model are working. Lastly, the result that is generated gives a clear view to the processing of the data and model.

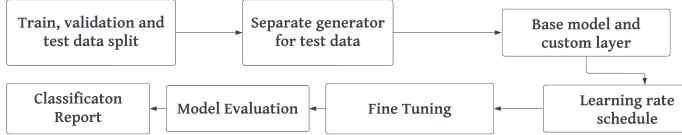


Fig. 2. The operational way of the models

V. RESULT ANALYSIS

The work precisely states that CNN models really have a great impact on the early direction and taking precaution beforehand. In this paper, comparing the crop disease and different methods to learn which works better is significant as shown in TABLE I.

A.

section*Comparison of Model Performance
label=0.

- 1) **Inceptionv3** achieved the highest accuracy among the models evaluated, with an accuracy of 83.63%.
- 2) **InceptionResNetV2** closely followed Inceptionv3 with an accuracy of 82.70%.
- 3) **ResNet50** showed a reasonable performance with an accuracy of 76.43%.
- 4) **VGG16** had the lowest accuracy among the models, with an accuracy of 66.95%.
- 5) In terms of precision, recall, and F1-score, **Inceptionv3** consistently outperformed the other models across all classes.
- 6) While **VGG16** had the lowest accuracy, it also had the lowest precision, recall, and F1-score across all classes.
- 7) **ResNet50** showed competitive performance but was outperformed by both Inceptionv3 and InceptionResNetV2 in terms of accuracy and precision.

- 8) Overall, **Inceptionv3** and **InceptionResNetV2** demonstrated superior performance compared to VGG16 and ResNet50 in classifying skin lesions.

TABLE I
MODEL ACCURACY

Model	Accuracy	Precision	Recall	F1 Score
InceptionResNetV2	82.70%	0.86	0.83	0.83
InceptionV3	83.63%	0.86	0.84	0.83
ResNet50	74.43 %	0.74	0.76	0.72
VGG16	66.95%	0.45	0.67	0.54

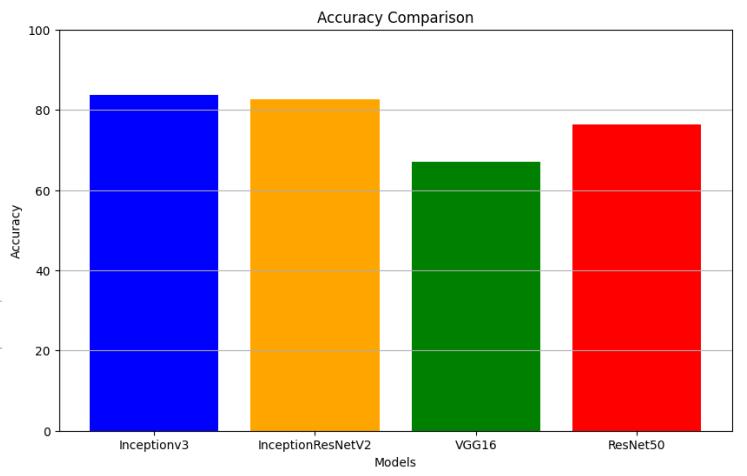


Fig. 3. Accuracy Comparison

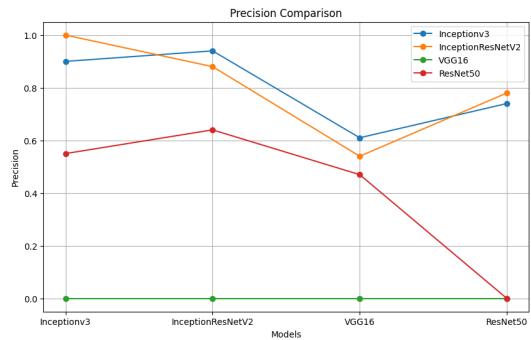


Fig. 4. APrecision Comparison

In this paper, each of the models run for the data set has set some precision, recall, f1-score, and support score. On this basis, some graphs are made to understand which one of the models worked better, as shown in Fig.3, Fig.5, Fig.7, and Fig.9.



Fig. 5. Inceptionv3 confusion matrix

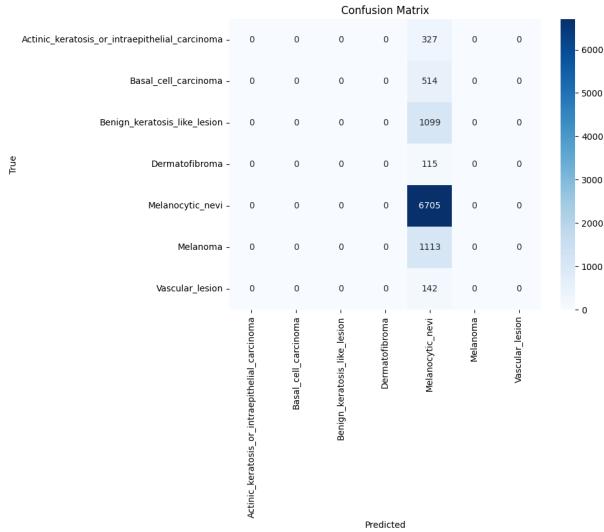


Fig. 8. VGG16 confusion matrix



Fig. 6. InceptionResNetV2 confusion matrix

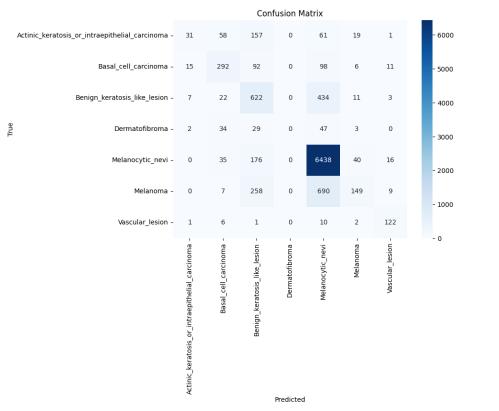


Fig. 7. ResNet50 confusion matrix

The InceptionResNetV2 has an accuracy of 82.70%, Inceptionv3 has an accuracy of 83.63%. Both of the models can have a great impact in the life span of the farmer as it could be helped in detecting the disease in an earlier manner. Moreover, the other models. ResNet50 has also done precisely well as this has accuracy of 0%. On the other hand, VGG16 has accuracy of 66.95% . In the Fig.7 a combined graph is given with the model to show which is performing better among all those. CNN significantly affects tomato plant leaf disease detection by improving result production.

In this paper, we have analyzed the interpretability of the InceptionResNetV2 image classifier on our hybrid tomato disease dataset using explainable AI. We have used Lime (Local Interpretable Model-agnostic Explanations), a very capable machine learning technique, to shed light on the process of decision-making of our InceptionResNetV2 because this model was able to classify new tomato leaf diseases with the highest accuracy. Lime analyzes the input image and observes the resultant changes by the specified model's output to accurately highlight the regions of the leaf image that contributed toward the decision-making of the specified classifier model. In this paper, we have shown two images side by side, the first one is an original image of a Tomato Early Blight disease, and the second one is the image output of lime. The green region on the second image highlights the regions based on which the InceptionResNetV2 classified the image to fall under Early Blight class, the more darkened region means this region had the highest influence for the decision making. We can conclude that InceptionResNetV2 was making all the decisions correctly based on our observation using lime output, as we can see that the majority of the leaf area was highlighted and very low outside area was highlighted by lime, this explains why this model was able to output results with highest accuracy. This technique enhances the transparency behind machine learning model decisions and also contributes

valuable insight into the robust architecture and classification capabilities of popular deep learning classification techniques.

VI. LIMITATION

The skin cancer and lesion classification are done using CNNs. But, CNNs require lots of labeled data to function properly. And CNNs may not work well if trained with little data or on very specific data. Due to the CNNs' architecture, they require large computational power and time, especially in more complex models. In addition to this, CNNs fail in retaining detailed spatial arrangement information, making the exact localization of the abnormality difficult. This can be addressed using pooling layers but not completely. Deeper layers can understand the context and patterns in the data, but they may miss out on the macro view. CNNs need a lot of labeled data to train effectively. If you do not have much data, the resulting model may be bad, or it might suffer from overfitting. CNNs need a training power, especially deep ones. Thus, it may take a long time and consume a lot of resources to train your model. Although convolutions allow CNNs to be invariant to translations, it is difficult for them to determine the exact locations of an object. Therefore, the model can focus on irrelevant areas and miss the main object of the image. CNNs are generally black boxes. We don't know why the model predicted a specific result.

For efficient training, Convolutional Neural Networks (CNNs) require a substantial amount of annotated data. Inadequate generalization of models or overfitting may occur due to insufficient data. Convolutional Neural Networks (CNNs) necessitate significant computational capacity for training, particularly for more complex architectures, resulting in time and resource-intensive processes. CNNs employ pooling layers to ensure invariance to translation, although this can lead to a reduction in spatial information. Consequently, they might encounter challenges when it comes to successfully finishing precise localization tasks. CNNs may disregard broader contextual information as they prioritize local information within their receptive fields.

CNNs require a lot of labeled data. Insufficient generalization of models or overfitting might result from insufficient data. CNNs require a lot of processing power to train, especially deeper structures, which makes the process time- and resource-consuming. CNNs use pooling layers to ensure translation invariance, however this might result in a loss of spatial information. They may thus find it difficult to complete accurate localization assignments. CNNs may overlook more comprehensive contextual information in their focus on local information inside receptive fields. Deeper networks could find it difficult to comprehend the global context. CNNs' robustness may be impacted by differences in input data, such as noise, occlusion, or changes in illumination and orientation. It might be difficult to comprehend why particular predictions are made by CNNs. They are frequently regarded as interpretability-deficient "black-box" models. Rather than looking at the dark side of the architecture, exploring the brightside would help to work efficiently.

VII. CONCLUSION

From the experiments in this study, we confirm that the preprocessing method is an essential step in its superiority in using Convolutional Neural Networks (CNNs) for skin cancer and skin lesion classification using the HAM10000 dataset. Some CNN architectures, such as InceptionResNetv2, Inceptionv3, ResNet50, and VGG16, have been used, and they are implemented correctly. Among these CNNs, the InceptionResNetv2 CNN demonstrating the best precision and recall values reflects the high performance of the model in accurately capturing skin lesion cases. However, in our future work, we are considering working on other CNN architectures such as vision transformers and modifying existing architectures to improve the performance of the CNN models. Additionally, the CNN models were also analyzed using Lime for model interpretability to provide an interpretation framework of these models. This data model will provide an understanding of decision-making in the CNN models. Our model helps improve skin cancer and skin lesion discrimination tools, and it holds a prime incentive for clinical and agriculture in cancer advance boundaries such as early stage detection and diagnostics.

REFERENCES

- [1] M. S. Hossain, S. N. Hassan, M. Al-Amin, M. N. Rahaman, R. Hossain, and M. I. Hossain. (2023). Kidney disease detection from ct images using a customized cnn model and deep learning.
- [2] Chang Jie, Chen Jiming, Shao Ying. (2023). A pyramid GNN model for CXR-based COVID-19 classification.
- [3] N. ul Huda, R. Amin, S. I. Gillani, M. Hussain, A. Ahmed, and H. Aldabbas. (2023) Skin cancer malignancy classification and segmentation using machine learning algorithms. Available: <https://www.sciencedirect.com/science/article/pii/S2772375523000084>
- [4] L. Tiwari, V. Awasthi, R. K. Patra, R. Miri, H. Raja, and N. Bhaskar, "Lung cancer detection using deep convolutional neural networks," in Data Engineering and Intelligent Computing: Proceedings of 5th ICICC 2021, Volume 1. Springer, 2022, pp. 373–385.
- [5] W. Sae-Lim, W. Wettayaprasit, and P. Aiyarak, "Convolutional neural networks using mobilenet for skin lesion classification," in 2019 16th international joint conference on computer science and software engineering (JCSSE). IEEE, 2019, pp. 242–24
- [6] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," nature, vol. 542, no. 7639, pp. 115–118, 2017.
- [7] Joshi, S. K. (2022). Comparison and Performance Evaluation Using Convolution Neural Network-Based Deep Learning Models for Skin Cancer Image Classification. In Advances in Deep Learning for Medical Image Analysis (pp. 53–68). CRC Press.