National College of Ireland

Skin Cancer Diagnosis using Image Augmentation and Machine learning approaches

MSc Research Project
Data Analytics

Pooja Rajesh Garje
Student ID: 22241001

School of Computing
National College of Ireland

Supervisor: Prof. Furqan Rustam

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Pooja Rajesh Garje
Student ID:	22241001
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Prof. Furqan Rustam
Submission Due Date:	16/09/2024
Project Title:	Skin Cancer Diagnosis using Image Augmentation and Machine learning approaches
Word Count:	6863
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Pooja Rajesh Garje
Date:	13th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Skin Cancer Diagnosis using Image Augmentation and Machine learning approaches

Pooja Rajesh Garje
22241001

Abstract

Skin cancer is a serious illness, and diagnosing it early ensures timely treatment. But the way doctors usually diagnose skin cancer can sometimes be subjective, which means there's a chance of misdiagnosis. Current machine learning approaches face significant challenges, such as handling imbalanced datasets where some skin cancer types are less represented. These methods also struggle with processing complex images, making it difficult to accurately classify lesions. To address these gaps, in this study we compare the effectiveness of Graph Neural Networks (GNN) against Convolutional Neural Networks (CNN), logistic regression and random forests methods in classifying skin cancer images. We tested these statistical models by applying them on Skin Cancer ISIC dataset using non-augmented and augmented datasets with specific focus on performance differences between these two types of datasets. These outcomes showed that the CNN model had 80% accuracy for non-augmented data sets performing better than other methods.The GNN based segmentation model realized potential in classifying images yet struggled to convert image to graphs thus resulting into only 48% accuracy on highest input resolution validation dataset. Other methods like Linear Regression and Random Forest did okay but didn't perform as well as CNN. The models performed well on the original dataset but struggled with the augmented dataset. Therefore, the future work will focus on investigating advanced data augmentation methods, such as (GAN) Generative Adversarial Networks and (SMOTE) Synthetic Minority Over Sampling Technique, while also utilizing larger and more varied datasets. Additionally, there is potential in merging the strengths of both CNN and GNN and incorporating other data, such as patient history, to enhance diagnostic accuracy.

Keywords: Skin cancer, image classification, Convolutional Neural Networks, Graph Neural Networks, data augmentation, Generative Adversarial Networks, Synthetic Minority Over-sampling Technique.

1 Introduction

The rate at which the cases of skin cancer are increasing among adults suggests that mitigation measures are required. The most common type of skin cancer are Non melanoma skin cancer(NMSC). They account for the majority of skin cancer cases across the world. However, as noted by (Elgamal; 2013) melanoma is less common yet the most deadly kind of skin cancer because of its tendency for metastasis and death when discovered later.The rates of morbidity and mortality associated with skin cancer have

increased significantly over time across the globe. For instance, alone in America there are about 60,000 new diagnoses of invasive melanoma each year with 8,000 resulting. According to (Hosny et al.; 2018) ,factors contributing to this rise include increased exposure to ultraviolet (UV) radiation from the sun, genetic predispositions, and various environmental influences. Early identification is crucial, as (Okuboyejo et al.; 2013) emphasize, since early diagnosis significantly increases the chances of full recovery; in cases of melanoma, the cure rate can reach up to 90% if treated early. However, the visual similarity between benign and malignant lesions often complicates diagnosis, making it challenging even for experienced dermatologists. Traditional diagnostic methods, which rely heavily on visual inspection and dermoscopy, are subjective and can lead to misdiagnosis and delayed treatment.

1.1 Research Motivation

To detect malignant skin cells, better diagnostic techniques are needed for a number of reasons.However, interpretation variations limit traditional approaches while some are manual hence taking time. Consequently, this limitation has necessitated more accurate and quicker diagnostic tools.As demonstrated by (Esteva et al.; 2017), CNNs have shown promise in medical imaging, particularly in skin cancer detection, as they can automatically identify patterns in images. However, CNNs often fail to capture the intricate spatial relationships and hierarchical structures within dermoscopic images which are crucial for accurate segmentation and classification.

1.2 Research Contributions

To counter these limitations, researchers such as (Scarselli et al.; 2008) have suggested the use of GNN, a recent method that uses graph based structures to represent spatial relationships. GNNs are capable of capturing complicated spatial and hierarchical interrelations in dermoscopic images.Our study seeks to evaluate the effectiveness of GNNs by comparing it with CNN and traditional approaches like logistic regression, random forest, KNN and decision trees. Besides, this paper is going to discuss clinical implications of our diagnostic procedure and recommend future developments within dermatology imaging research targeting skin cancer. Our goal in carrying out this research has been to find better diagnostic tools for dermatologists when dealing with skin cancer patients than what they have been using until now.

1.3 Research Questions and Objectives

How can GNN improve skin cancer image classification compared to CNN and traditional machine learning models, and what are their advantages and limitations in clinical diagnosis?

Objectives:

1. Develop a GNN based segmentation method for skin cancer image classification, incorporating advanced image augmentation techniques to improve model performance.
2. Compare the performance of GNN based segmentation model with CNN and traditional machine learning models such as Logistic Regression, Random Forest, KNN, and Decision Trees.

3. Evaluate the advantages and limitations of using GNNs for skin cancer diagnosis in comparison to conventional image analysis methods.

1.4 Structure of paper

This report will be broken down into seven main parts. research is introduced in section 1, which focuses on the motivation and objectives. section 2 reviews literature, thereby situating it within a broader context and exposing its gaps. Methodology is detailed in section 3, outlining the data processing, model implementation, and evaluation metrics. The design specifications of implemented models are overviewed in section 4. CNN, GNN, and traditional machine learning models were all put into practice in section 5. Results are presented by section 6 together with comparisons with existing studies' performance. Finally, the seventh part concludes on research findings and proposes further work.

2 Related Work

The majority of skin cancer cases, a common form of the disease, are typically diagnosed through visual examination. This process starts with a clinical assessment, followed by magnification using a device called a dermoscope, then a biopsy analysis. Classifying skin lesions automatically from their images still remains an unachievable goal as they are subtly and finely heterogeneous.

In their study,(Esteva et al.; 2017)developed a CNN model for classifying dermatological conditions using a hierarchical taxonomy of 2,032 diseases grouped as benign, malignant or non-neoplastic lesions. The model was pretrained on ImageNet and fine-tuned with dermatological images from well known datasets such as ISIC Dermoscopic Archive, Edinburgh Dermofit Library and Stanford Hospital that were also used for training the model, which is based on Google Inception v3 architecture. In both nine way and three way classifications the CNN had higher accuracy than 21 dermatologists due its diagnostic prowess reflected through sensitivity-specificity curves, confusion matrices as well as saliency maps. Despite these achievements it had some limitations including possible labeling inaccuracies associated with non-biopsy proven images and test set variability. This highlights the critical role of high-quality, biopsy confirmed images in training robust models.(Medhat et al.; 2022) proposed a methodology that compared the effectiveness of different CNN architecture like AlexNet, MobileNetV2, and ResNet50 in diagnosing skin cancer using smartphone images.They used 2,298 images in their experiment, hence it was found that AlexNet had the highest accuracy of 0.99. Notably, AlexNet consumed high amounts of power and took long to learn when applied to the dataset as opposed to ResNet50 which learned faster than others. The study highlighted the trade-offs between computational demands and diagnostic performance, emphasizing the need for balanced datasets to avoid overfitting.

In their study (Aljohani and Turki; 2022) proposed a deep learning approach, which employs CNN architectures such as DenseNet201, MobileNetV2, ResNet50V2, Xception, VGG16, VGG19 and GoogleNet(Inception v1) to melanoma classification. The best performer was noted to be GoogleNet with a score of 0.76; whereas the scores for DenseNet201 was 0.74, followed by ResNet50V2 (0.73). Nevertheless, there were several limitations such as small size of data set and its quality restriction including no transparency among others impacting interpretability through black box nature of the CNNs. They also

advocated for future validation on more diverse datasets. In research done by (Xin et al.; 2022) they developed a Vision Transformer (VIT) model for skin cancer classification integrating multiscale feature extraction with contrastive learning. However, when tested on HAM10000 this approach attained an AUC of 0.987 but it lacked behind because it is very slow when processing on high resolution images and has poor interpretability due to complex decision interpretation issues . It is important to note that this investigation shows why traditional CNNs can be outperformed by VITs especially when representing long range dependencies inside images.

The study by (Soudani and Barhoumi; 2019) proposed a segmentation recommender system combining crowd sourcing and transfer learning to enhance skin lesion extraction. The study utilized ISIC2017 dataset.The accuracy for the VGG16 based model was 0.76, whereas the accuracy for the ResNet50 based model was 0.73. The ResNet50 model recorded the best sensitivity, but both models markedly increased the Dice coefficient and Jaccard index score.The study emphasized the potential benefits of integrating crowd sourcing with deep learning, but it also pointed out drawbacks, like the need for high quality expert annotations and dataset size.To address these issues, it is recommended that future studies investigate the combination of more sophisticated data labeling method with semisupervised learning strategies.

Deep transfer learning was used to evaluate skin cancer in (AL-SAEDİ and Savaş; 2022) using DenseNet, Xception, InceptionResNetV2, ResNet50, MobileNetV2 and EfficientNet models on the ISIC dataset. Among all other predictors, DenseNet121 showed the highest accuracy of 99.6% with better precision, recall, F1score and specificity over others. It highlighted the difficulties like class imbalance and computationally expensive operations.(Ashfaq and Ahmad; 2023) researched on the use of deep learning methods like Vision Transformers for detecting melanoma. For instance, they tuned HAM10000 dataset on several architectures such as VGG16, ResNet50, InceptionResNetv2 and EfficientNetB3. From a comparison of their accuracies and F1 scores it was discovered that EfficientNetB3 was performing well compared to other architectures. Nonetheless; Vision Transformers had low performance indicating that it is not possible to directly use transformer architectures without much pretraining data primarily designed for medical imaging purposes. Transformer models have been widely used in medical imaging but more diverse and bigger training sets are needed to fully exploit them.

Similarly, the paper by (Aldwgeri and Abubacker; 2019) classified the skin lesions using an ensemble of deep CNNs on dermoscopy images using pre-trained models such as VGGNet, ResNet50, InceptionV3, Xception, DenseNet121 and transfer learning. The study attempts to address class imbalance in 10015 images over seven classes from ISIC 2018 challenge dataset through techniques like data augmentation and weight balancing. The ensemble model achieved a balanced accuracy of 80% and a mean AUC of 0.89, outperforming individual models. This study suggests incorporating demographic data to improve accuracy and generalize the diagnostic network. (Almaraz-Damian et al.; 2020) suggested the use of CAD system that combines handcrafted features with deep learning features using Mutual Information measures. MobileNet v2 architecture along with hand crafted features achieved highest accuracy of 92.4% IBA was reported to be equal to 0.80 in the ISIC 2018 dataset, however, it achieved a poor lesion segmentation accuracy with computational complexity as well.

In the work of (Zia Ur Rehman et al.; 2022) propose deep learning methods like

MobileNetV2, DenseNet201 to improve the classification and localization of skin cancer. The dataset was collected from Kaggle and the results indicate that the accuracy of MobileNetV2 was 0.90 and that of DenseNet201 was 0.94. Both models demonstrated, sensitivity, specificity, precision, and F1 scores validated through GradCAM were high. These main challenges, however, were related to scalability and interpretability which stressed the dependence of such robust and generalizable models on the clinical practice. In the work of (Ameri; 2020) was adapted AlexNet in transfer learning to classify the skin lesions as malignant or benign. Performance was affected by limitations such as dataset imbalance and using only a small portion of images available in the dataset.

2.1 Research limitation and gaps:

The literature draws attention to several limitations and gaps in current research based on classification of skin cancer using machine learning models. An important limitation is addressed by (Esteva et al.; 2017) that there are differences in image quality and incorrect labels because they come from images that were not confirmed by biopsy. Another challenge lies in balancing computational demands and diagnostic performance, as demonstrated in the comparisons between AlexNet and ResNet50 by (Medhat et al.; 2022).(AL-SAEDİ and Savaş; 2022) further emphasized that data imbalance is a major factor affecting model performance and generalization. In order to boost its robustness and accuracy, however, the models would have been better served by more diverse training datasets of larger sizes along with advanced techniques for data augmentation especially when it comes to classes which are underrepresented.

While CNNs and transformer based models show promise, challenges such as data imbalance, data quality, computational demands and interpretability need addressing. This study takes this opportunity to employ models such as CNN, GNN and various machine learning models to further improve skin cancer disease prediction accuracy.

3 Methodology

The proposed methodology comprises of several stages, each stage contributes to the overall effectiveness of the classification process. As per the Figure 1, the data is extracted from kaggle and stored locally. This original dataset is used to generate a separate augmented dataset. Afterward, these datasets are partitioned into training and validation sets. Various machine learning models as well as deep learning models such as GNN, CNN, LR, Random Forest, KNN and Decision Trees are trained on training data. Then the models are tested on validation data. Model performance is measured using various evaluation metrics including accuracy, precision, recall and F1 score.

3.1 Dataset Description

This study was carried out on the Skin Cancer ISIC dataset (*SkinCancerISIC*; n.d.), provided by the International Skin Imaging Collaboration (ISIC). It is ideal as it provides a wide range of images for skin lesions necessary to create strong machine learning models. This dataset, which can be found on Kaggle, includes nine different image categories: Actinic Keratosis,, Squamous Cell Carcinoma, Dermatofibroma, Melanoma, Nevus, Basal Cell Carcinoma, Seborrheic Keratosis, Pigmented Benign Keratosis and Vascular Lesion.

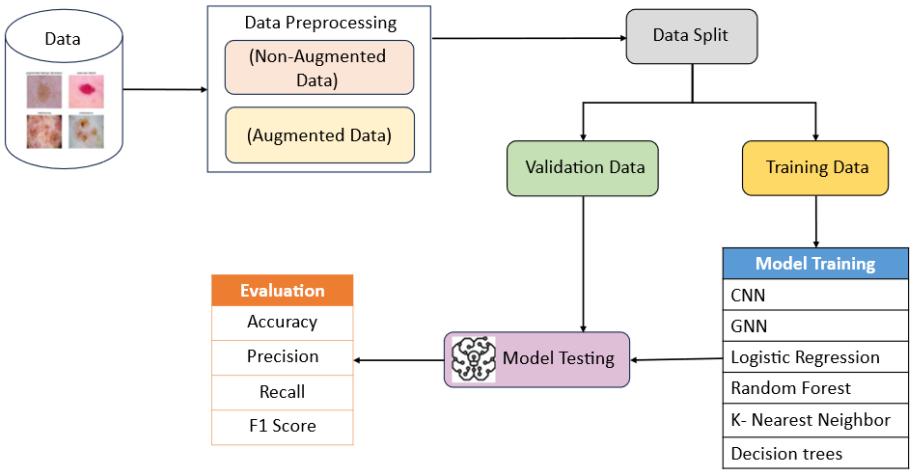

Figure 1: Proposed Methodology Architecture

There are a total of 2,357 images that include 2,239 pictures for training and 118 ones for testing it offers a vast source in relation to precisely identifying different variations of cancerous tumors affecting the skins. The inclusion of both malignant and benign oncological diseases meticulously handpicked from ISIC archive represents a wide variety sample that improves the model's ability to be used for other practical cases thus giving it a better chance of being applied in real-life situations.

3.2 Data Analysis

The data from this investigation revealed that there is a high imbalance in distribution among classes as shown in Figure 2. Some classes like melanoma and pigmented benign keratosis have much more images compared with dermatofibroma and seborrheic keratosis amongst others. Such discrepancies could be detrimental to model training because they introduce bias towards the majority classes when making predictions.

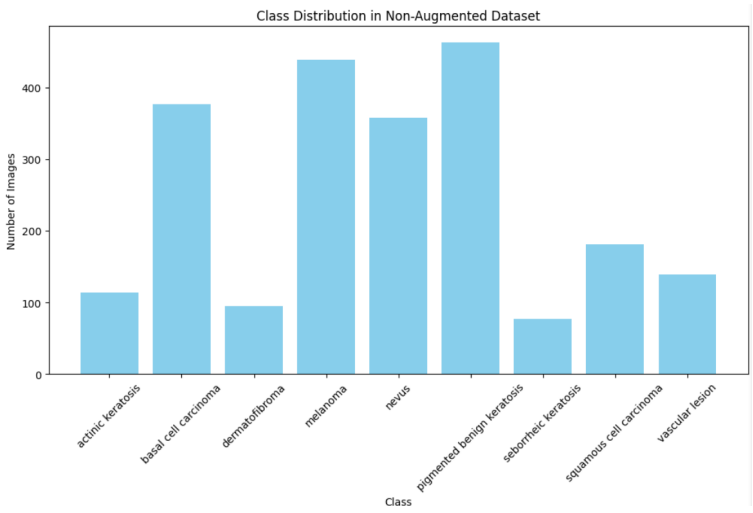

Figure 2: Class Distribution before augmentation

The main issue now becomes how to handle a situation where other categories are not adequately learned by machine learning because of their small population which will then lead to biased predictions from machine learning models. As stated in the study by (Ameri; 2020) it is important to address this using data augmentation as it helps in building robust and accurate classification models.

3.3 Data Augmentation

The model's generalizability was improved by the data augmentation technique because it balanced the class distribution within the dataset as explained by (AL-SAEDİ and Savaş; 2022).The process involved applying various transformations to the images in classes with less than 400 images.The transformations generated multiple variations of each image, thereby increasing the number of samples in underrepresented classes as shown in Figure 3.The augmented images were integrated with the original dataset, ensuring each class had atleast 400 images. This process effectively mitigated class imbalance and provided a more robust training dataset for the models.

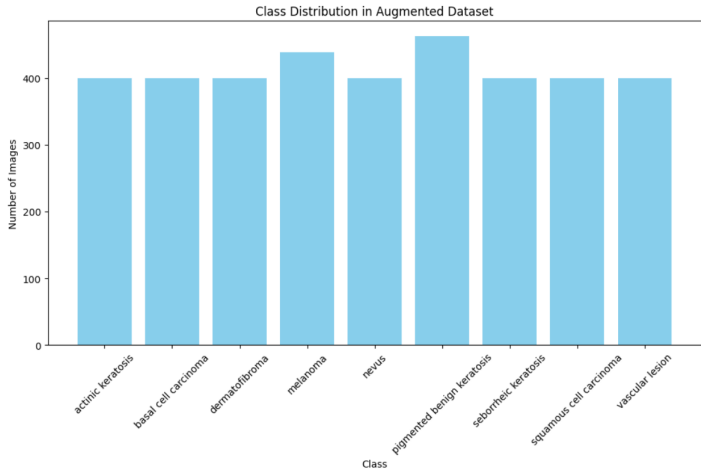

Figure 3: Class Distribution after augmentation

3.4 Data Split

The dataset was divided into training and validation sets to ensure that the model could be properly learned and tested.The data was randomly split into 80:20 ratio, where 80% of it is a training set 20% is used as a validation set.The training set was used to train the models while the validation set was utilized in establishing model generalization performance.

3.5 Model Training

In model training phase the different machine learning and deep learning are trained on the preprocessed training data.This stage involves feeding the models input images, adjusting weights and assessing performance on the training set. The CNN model is known for its ability to capture spatial patterns effectively by automatically learning

and extracting hierarchical features through convolutional, pooling, and fully connected layers. This study by (Cullell-Dalmau et al.; 2021) highlights the CNN's effectiveness in learning intricate features from images, which significantly contributes to its performance in classification tasks. The GNN model transforms images into graph representations such as superpixels as nodes and edges learn complex spatial dependencies through message passing techniques. This method (Wu et al.; 2022) helps in capturing relational and structural information which is crucial for accurate classification.The GNN model shows great potential in managing complex relationships between image elements, even though converting images into graphs presents some difficulties. Random Forests reduce overfitting by constructing multiple decision trees, enhancing prediction accuracy (Dinesh et al.; 2024). KNN works by comparing input samples with its 'k' nearest neighbors, assigning the most common class label. Decision Trees split data into subsets based on feature values, resulting in an interpretable tree like structure that can be pruned to minimize overfitting.

3.6 Model Testing

In the model testing phase, after the hyperparameter tuning the trained models are also evaluated using a separate validation dataset to measure their performance on unseen data. This stage involves giving validation images to these machine learning algorithms as well as comparing predictions and actual labels. This process encompasses evaluating several indices such as accuracy, precision, recall,and F1 score which measure how good are these models. After analyzing the results, the best performing model, is identified and this model is then tested on a test dataset which ensures it's robustness.

3.7 Model Evaluation

To evaluate the models, in the project the following metrics are considered:

3.7.1 Accuracy

Accuracy is the percentage of correctly grouped instances from the overall number of instances.
The accuracy of a model is calculated using the below formula:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

Where:

- "Number of Correct Predictions (TP+TN)" is the count of predictions that match the true labels.
- "Total Number of Predictions (TP+TN+FP+FN)" is the total count of predictions made by the model.

Here, TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

3.7.2 Precision

As per (Zia Ur Rehman et al.; 2022) precision is the ratio of correctly predicted positive observations to the total predicted positives . High precision implies low rate of false positives which is important in medical diagnosis as these will indicate healthy patients as having cancer.

3.7.3 Recall

Recall is the ratio of correctly predicted positive observations to all observations in the actual class. High recall is essential for identifying most of the true positive cases, crucial for ensuring cancer cases are not missed as stated in (Ameri; 2020).

3.7.4 F1-score

The F1-score is the mean of precision and recall. As stated in the study by (Zia Ur Rehman et al.; 2022),it provides a balance between precision and recall, especially useful when the class distribution is imbalanced.

3.7.5 Confusion Matrix

The equation used to compute the confusion matrix is displayed in Figure 4.

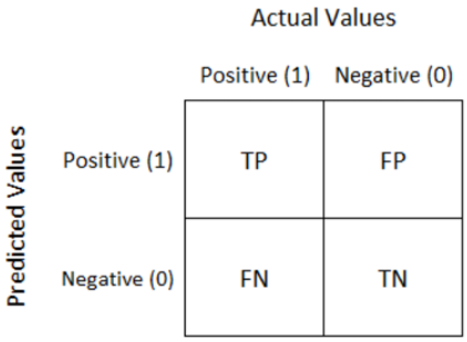

Figure 4: Confusion matrix for classification model

By offering a thorough breakdown of accurate and inaccurate classifications, the confusion matrix makes it possible to analyze model performance more precisely across many classes.

4 Design Specification

This project is implemented on a system with 12th Gen Intel Core i5 processor. It runs on a 64 bit operating system, with 16 GB of RAM, and Intel iRIS graphics. The models were created and tested on an Anaconda environment to ensure consistent package management and dependencies. Jupyter Notebooks was used for interactive coding, testing and visualizing outcomes.

To ensure a clearer understanding of the models, the study initially set out to prepare and augment the image dataset for the models. Our work began with building CNN

employing TensorFlow and Keras, further optimizing the models with the Adam optimizer and crossentropy loss to enhance the model's performance while minimizing errors. The next step focused on employing GNN using PyTorch and Torch Geometric in order to analyze images differently than before. We also compared the results with traditional machine learning techniques. The goal of this approach was to image skin cancer accurately by classifying the different images into the best performing models.

4.1 Convolutional Neural Network (CNN)

The CNN model comprises of two parts where the first part contains the CNN layers for extracting the features and the second part contains the fully connected layers. In our proposed model, the first part of CNN comprises three convolutional layers, each subsequent to ReLU activation function, a max pooling layer and dropout layer. The convolutional layers are configured with 64, 128, and 256 filters, respectively, each with a 3x3 kernel size, as described by (Reshi et al.; 2021), which aids in extracting features from input images.The max pooling layers, each with a pool size of 2x2, are applied after each convolutional layer to reduce the spatial dimensions of the feature maps, as highlighted by (Rustam et al.; 2022). Dropout layers with a dropout rate of 0.2 are included after each max-pooling layer to prevent overfitting.

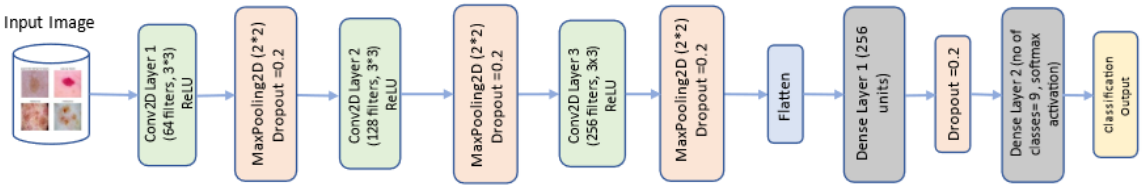

Figure 5: The CNN model architecture

The second part of the CNN consists of a flatten layer followed by two fully connected layers with 256 neurons each, and a dropout layer with a 0.2 dropout rate. Finally, there's an output softmax layer with nine neurons fitting the nine classes of the skin lesions illustrated in Figure 5. This way, the model is capable of understanding complex feature relationships and can produce the correct output.

4.2 Graph Neural Network (GNN)

In this research, to classify skin cancer and its states the GNN model was applied which included a layer called SimpleMessagePassing allowing to average the features of the node and thereby change and pass it. The structure of the GNN is represented in the SimpleGNNModel with two layers of message passing 64 and 128 channels of output respectively. As stated in the study by (Scarselli et al.; 2008) each of these layer is followed by ReLU activation functions to introduce non linearity. A global mean pooling layer after message passing aggregates node features into a graph level representation. It also serves as an indication that the model is trained on labeled data only(the training

set). This is followed by a fully connected final layer whose output size equals the number of classes in order to predict class labels. By doing this, the GNN can capture fine-grained image patterns through graph representations and learn from them.

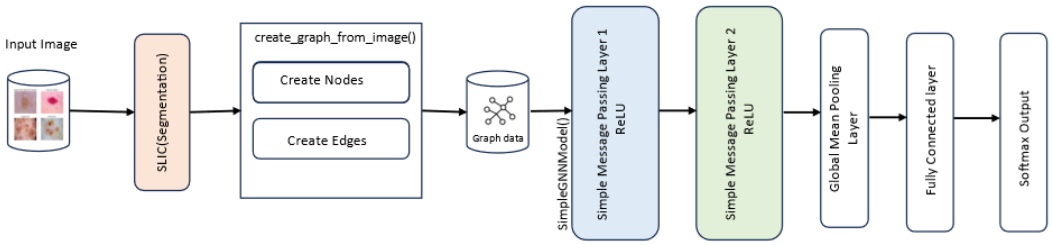

Figure 6: GNN model architecture

As shown in the above Figure 6 data preprocessing involves converting images into graph structures by applying the SLIC algorithm for superpixel segmentation, where each superpixel is represented by its mean color as a node feature and edges are based on adjacency in space. To balance the training dataset especially for less frequent cases, a variety of data augmentation methods are used as mentioned in subsection 5.1. The learning rate is set to 0.001 and Adam optimizer is used during training along with crossentropy loss function. Training lasts for 50 epochs followed by evaluation of model performance using accuracy, precision, recall, F1 score and confusion matrix metrics. Such detailed strategy combining GNNs relational learning with data augmentation increases sensitivity and robustness of skin cancer classification model.

5 Implementation

This section describes the specific procedures and techniques used in creating and implementing the suggested diagnostic model for the identification of skin cancer.

5.1 Implementation of Data Augmentation

The training dataset's class imbalance was addressed by applying the data augmentation method, which made sure each class contained at least 400 photos. Many transformations, such as rotation, horizontal flipping, zooming, contrast, brightness, color shift, and random distortion, were applied using the Augmentor package, with the values listed in Table 1.These techniques generated new images to increase the dataset size for under-represented classes. The augmented dataset was then reloaded, resized to required input resolutions, and split into training and validation sets.

5.2 Implementation of CNN

The aim was to find the smallest size an image could be while preserving its classification accuracy.Different CNN models were implemented to find the appropriate image size as input for the final model. Each model had some convolutional layers followed by max-pooling layers that reduced spatial dimensions and fully connected dense layers for final

Table 1: Data Augmentation Parameters

Augmentation Technique	Probability	Parameters
Rotation	0.7	Max Left Rotation: 10°, Max Right Rotation: 10°
Flip (Left-Right)	0.5	NA
Zoom (Random)	0.5	Percentage Area: 80\%
Contrast Adjustment	0.5	Min Factor: 0.5, Max Factor: 2.0
Brightness Adjustment	0.5	Min Factor: 0.5, Max Factor: 1.5
Color Shift	0.5	Min Factor: 0.5, Max Factor: 1.5
Random Distortion	0.5	Grid Width: 4, Grid Height: 4, Magnitude: 8

classification.We also utilized some python packages in this programming to make it easy during development. The main frameworks used to build network architecture and optimize training process are TensorFlow and Keras libraries implemented for creating and training CNN models as shown in (Reshi et al.; 2021). The tensorflow.keras module turned out to be very helpful while creating convolutional, pooling, dense layers, Adam optimizer along with SparseCategoricalCrossentropy loss function were additionally employed to improve its performance.

The matplotlib.pyplot package was used to generate visualizations, including confusion matrices and training and validation curves, illustrating the models' performance over time. Additionally, numpy was essential for numerical operations and data manipulation, ensuring efficient preprocessing of the image datasets. The pathlib and os modules were

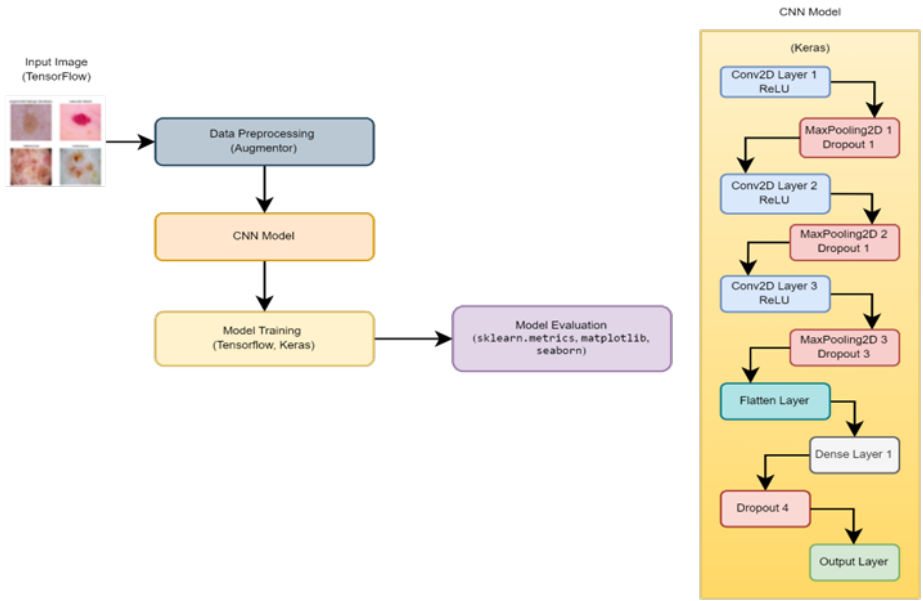

Figure 7: Implementation of CNN model

used for handling dataset and directory structure that will allow seamless navigation and organization of image files. As mentioned in the subsection 5.1, data augmentation is a critical step to improve model generalization. The sklearn.metrics module contained

functions such as classification report and confusion matrix which provided more details about the model's accuracy regarding classification and performance metrics per class. Besides, through seaborn heatmaps were created for confusion matrices making it easier to understand how well the model predicts.

The model building began with data preparation, where the images were resized to required resolutions and augmented to increase training set. Each CNN model was then constructed with a unique architecture suitable for input image size, having convolutional layers with ReLU activation function, max-pooling layers for down-sampling and dense layers with softmax activation functions for classification. Initially, models were trained using a batch size of 16 but later on it was increased to 32 which resulted in better model performance and faster convergence by stabilizing gradient updates.The initial learning rate was set to 0.001, with cross-entropy loss guiding the optimization process through the Adam optimizer, as described by (IEEE; 2018). The model was trained for 50 epochs.The post-training evaluation involved creating confusion matrices and classification reports for each of the models. The training and validation accuracy and loss curves show the learning dynamics as well as model convergence.

5.3 Implementation of GNN

In this study we created a GNN based image segmentation model for image classification. It is an approach to utilize GNNs in detecting skin cancer and constructing several technical layers and tools with a step by step guide. First, OpenCV is used to preprocess images loaded and resized to the same resolution using torchvision transforms. Rescaling image pixel values into some standard range involves normalizing all transform functions in Torchvision according to Raju et al. (2020). Specifically, each RGB channel uses mean values [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225], which are applied for normalizing the images.To achieve this purpose, it ensures that the input data is standardized thereby improving the convergence speed and stability during training of the GNN.This process employs simple linear iterative clustering (SLIC), implemented by scikit-image (Achanta et al.; 2010), which groups similar colored pixels into superpixels thus forming nodes in the graph.

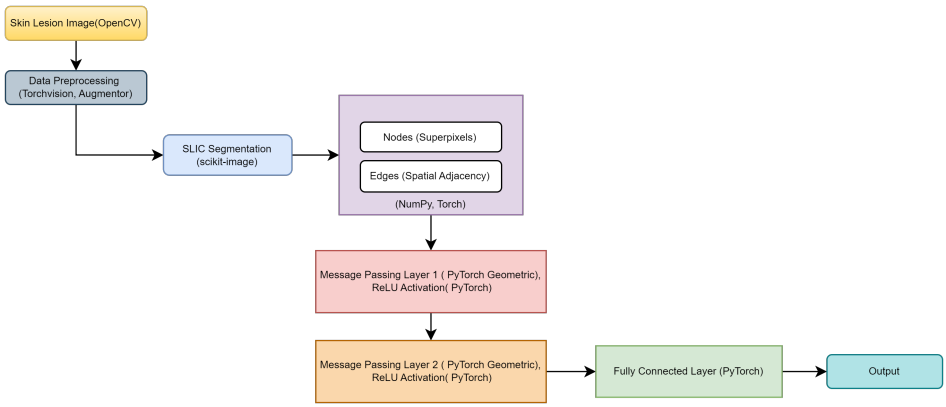

Figure 8: Implementation of GNN model

The Figure 8 shows how PyTorch and Torch Geometric are employed in constructing

the graphs. The node features are created by taking the average color of all superpixels while edges are determined through spatial proximity forming graph's adjacency matrix. There is GNN model architecture consisting of SimpleMessagePassing layers where node features undergo linear transformation followed by mean aggregation. After message passing comes ReLU activation which introduces non linearity. The node features are then aggregated to create a graph level feature vector using Global Mean Pooling that finally maps them to output classes via a fully connected Linear Layer.Training and optimization are handled by PyTorch, utilizing Cross-Entropy Loss to measure prediction accuracy and the Adam Optimizer for parameter updates, set with a learning rate of 0.001. The model is trained over 50 epochs, monitoring performance through training and validation accuracies. For model evaluation, scikit learn metrics are calculated, with visualizations provided by Matplotlib and Seaborn. This comprehensive approach leverages advanced image processing, graph based neural networks, and standard machine learning evaluation tools to implement and assess the GNN model for skin cancer classification.

5.4 Implementation of Machine Learning Models

This study aimed at implementing machine learning models for the purpose of benchmarking their performance against deep learning models. The implementation began by transforming raw image data into an appropriate format. All these four models were trained on training set and further evaluated on the validation set. In order to ensure convergence, maximum 1000 iterations were used by the LR model with lbfgs solver. We selected Python programming language due to its libraries such as numpy, cv2 (OpenCV), pandas, sklearn, seaborn and matplotlib which facilitate efficient handling of data and visualization. As shown in Table 2 a thorough analysis of all the machine learning models was performed to analyze the impact of input image variance and data augmentation on performance and accuracy.

Table 2: Hyperparameters for machine learning models

Model Name	Hyperparameters
LR	max_iter=1000
RF	n_estimators=300, max_depth=300
KNN	default
Decision Tree	default

The first step involved preparing the data for model training. Images were sourced from two two different directories, augmented and non-augmented . All these directories were already divided into subdirectories with various skin cancer categories. From these subfolders, we extracted the labels and paths of the images. In each resolution, a generator function was employed to pre-process image batches. A batch consisted of an image read with cv2, converted to RGB format, resized by the target resolution and flattened as a feature vector by numpy. Then they were used as datasets for modeling with their respective tags.

6 Results

In this paper, the results section offers a detailed examination of how each model performed. The detailed classification reports and confusion matrix are printed to examine the results and present the findings. They involve measures like accuracy, precision, recall and F1-score which are used to evaluate effectiveness of models for skin lesion classification.

6.1 Performance of CNN model

The results are discussed in terms of confusion matrices, training and validation accuracy as well as loss curves for augmented and non-augmented datasets.

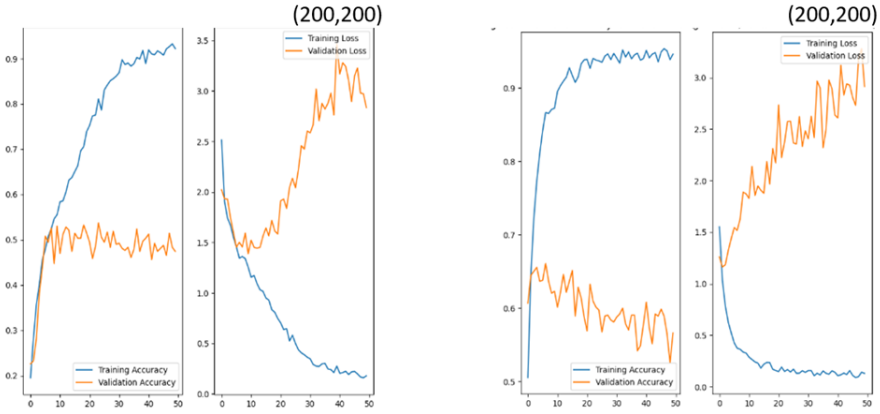

Figure 9: CNN model Training vs Validation Accuracy/Loss (200,200)

As per the Figure 9, The training accuracy of the original dataset increases steadily to exceed 90%, but validation accuracy gets stuck between 50-60% with a large value of validation loss that indicates poor generalization.In the augmented dataset, training accuracy also remains high, but validation accuracy fluctuates and validation loss stays high, suggesting that augmentation increases task complexity and challenges model accuracy and loss reduction. Our CNN models have a consistent accuracy of 84% for original dataset over different resolutions.

As per the result metrics of CNN model shown in Figure 10, The highest precision and recall were recorded by "basal cell carcinoma" at 100x100 pixels.Also, "dermatofibroma" and "squamous cell carcinoma" performed well; whereas "actinic keratosis" and "seborrheic keratosis" had low precision and recall. Augmented datasets showed overall decreased accuracies (54-57%), with "vascular lesions" having much higher precision and recall values than any other type of lesion.Moderate gains were observed despite an increase in resolution, indicating that image resolution alone is less important than model architecture and training strategy for instance. Instances of low sensitivity are due to augmented dataset's lower precision rates as noted for melanoma and actinic keratosis.

The best performance is observed on 200x200 resolution since it gave high accuracy rate and most stable loss in comparison with other inputs. In terms of input size, this resolution was selected as the optimal one for our CNN model.Augmented dataset showed a significant improvement in predicting skin cancer as compared to non-augmented one. It predicted 75 out of 84 cases correctly. Basal cell carcinoma and melanoma has high

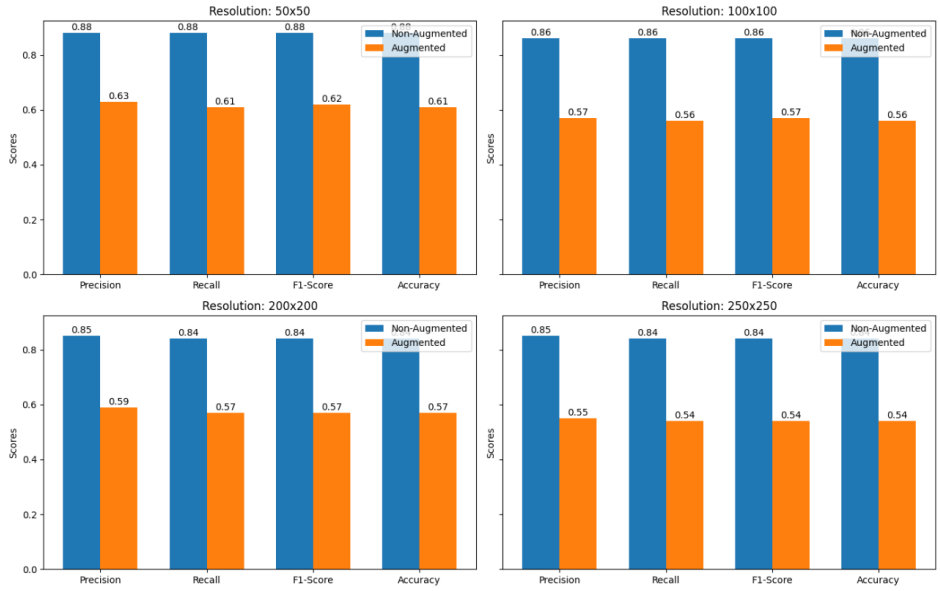

Figure 10: Performance Metrics (Weighted avg) by resolution

precision and recall values that demonstrate good model performance on early recognition. It did not perform accurately with seborrheic keratosis and actinic keratosis thereby lowering recall as well as precision due to misclassification. Also, these results showed that

Non-augmented dataset results

Classification Report (Non-Augmented Dataset):

	precision	recall	f1-score	support
actinic keratosis	0.41	0.32	0.36	22
basal cell carcinoma	0.96	0.89	0.93	84
dermatofibroma	0.87	0.87	0.87	23
melanoma	0.83	0.83	0.83	82
nevus	0.78	0.84	0.81	64
pigmented benign keratosis	0.94	0.92	0.93	83
seborrheic keratosis	0.38	0.53	0.44	17
squamous cell carcinoma	0.88	0.90	0.89	42
vascular lesion	0.97	0.97	0.97	30
accuracy			0.84	447
macro avg	0.78	0.79	0.78	447
weighted avg	0.85	0.84	0.84	447

Augmented dataset results

Classification Report (Augmented Dataset):

	precision	recall	f1-score	support
actinic keratosis	0.33	0.36	0.34	75
basal cell carcinoma	0.63	0.55	0.59	87
dermatofibroma	0.70	0.56	0.62	88
melanoma	0.36	0.48	0.41	73
nevus	0.59	0.57	0.58	82
pigmented benign keratosis	0.73	0.66	0.69	101
seborrheic keratosis	0.44	0.59	0.51	75
squamous cell carcinoma	0.46	0.41	0.43	71
vascular lesion	0.89	0.83	0.86	88
accuracy			0.57	740
macro avg	0.57	0.56	0.56	740
weighted avg	0.59	0.57	0.57	740

Figure 11: Classification report for CNN model with input size(200,200)

the number of false positives increased but there was no change regarding the correctness of predictions.As per the results shown in Figure 11 the model's performance on basal cell carcinoma dropped with only 48 correct predictions out of 87, highlighting the challenge with augmented data.

Overall, increasing image resolution improved training accuracy but suggested significant overfitting at lower resolutions. Data augmentation yielded mixed results, with some class improvements but limited overall performance impact.

6.2 Performance of GNN model

The GNN model's performance is evaluated, and a thorough analysis of the findings is provided. Since a validation accuracy of 48.21% was attained at image resolution of (250,250) for original dataset.The findings in Table 3 suggest that higher resolution typically improves model performance.Basal cell carcinoma was the skin lesion with the highest recall value; however, the classifications of dermatofibroma, seborrheic keratosis, and squamous cell carcinoma were difficult to categorize and had low recall and precision.

Table 3: Performance metrics of GNN model

Models	Image Resolution	Non- Augmented Dataset				Augmented Dataset			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
GNN	(50,50)	0.50	0.48	0.50	0.46	0.40	0.41	0.40	0.39
	(100,100)	0.47	0.45	0.47	0.42	0.37	0.38	0.37	0.36
	(200,200)	0.47	0.49	0.47	0.43	0.38	0.38	0.38	0.36
	(250,250)	0.48	0.44	0.48	0.45	0.39	0.40	0.39	0.37

From the above Table 3, it can be concluded that the accuracy of augmented datasets was generally lower than that of the original ones across all resolutions with best at (50,50) resolution being 40%. Particularly, higher resolutions improved precision, recall, and F1-scores among the original datasets. However, as expected there was no improvement in

Non-augmented dataset

Classification Report:

	precision	recall	f1-score	support
0	0.11	0.04	0.06	23
1	0.38	0.73	0.50	75
2	0.00	0.00	0.00	19
3	0.61	0.66	0.63	88
4	0.53	0.37	0.43	71
5	0.55	0.62	0.59	93
6	0.00	0.00	0.00	15
7	0.24	0.14	0.18	36
8	0.57	0.46	0.51	28
accuracy			0.48	448
macro avg	0.33	0.34	0.32	448
weighted avg	0.44	0.48	0.45	448

Augmented dataset

Classification Report:

	precision	recall	f1-score	support
0	0.30	0.42	0.35	80
1	0.28	0.57	0.37	80
2	0.31	0.06	0.10	80
3	0.45	0.51	0.48	88
4	0.62	0.54	0.58	80
5	0.35	0.42	0.39	92
6	0.60	0.39	0.47	80
7	0.29	0.09	0.13	80
8	0.43	0.49	0.46	80
accuracy			0.39	740
macro avg	0.40	0.39	0.37	740
weighted avg	0.40	0.39	0.37	740

Figure 12: Classification report for GNN model with input size (250,250)

the performance.This shows the need for more future experiments with augmentation.

6.3 Performance of Machine Learning Models

In this study, various machine learning models were utilized to analyze the performance of CNN and GNN models. The performance parameters and accuracies are compared for an in depth analysis of all machine learning models.

The results in Table 4 shows that the LR model accuracy for actual dataset ranged between 0.44 to 0.48 and augmented dataset was between 0.37 to 0.40.The KNN model had moderate performance where accuracies on real dataset ranges from 0.45 to 0.46 over all resolutions while precision, recall as well as F1score were a bit lower suggesting

Table 4: Performance metrics of machine learning models

Models	Image Resolution	Non- Augmented Dataset				Augmented Dataset			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Logistic Regression	(50,50)	0.45	0.45	0.45	0.45	0.37	0.37	0.37	0.37
	(100,100)	0.44	0.45	0.44	0.44	0.39	0.38	0.39	0.38
	(200,200)	0.44	0.46	0.44	0.45	0.37	0.37	0.37	0.37
	(250,250)	0.48	0.48	0.48	0.48	0.40	0.39	0.40	0.39
Random Forest	(50,50)	0.49	0.48	0.49	0.47	0.46	0.46	0.46	0.45
	(100,100)	0.49	0.48	0.49	0.47	0.44	0.44	0.44	0.44
	(200,200)	0.50	0.48	0.50	0.48	0.46	0.47	0.46	0.46
	(250,250)	0.49	0.48	0.49	0.47	0.46	0.46	0.46	0.46
KNN	(50,50)	0.45	0.44	0.45	0.42	0.35	0.36	0.35	0.33
	(100,100)	0.45	0.43	0.45	0.42	0.34	0.35	0.34	0.33
	(200,200)	0.46	0.43	0.46	0.43	0.34	0.35	0.34	0.33
	(250,250)	0.46	0.43	0.46	0.43	0.34	0.35	0.34	0.33
Decision Tree	(50,50)	0.32	0.34	0.32	0.33	0.26	0.26	0.26	0.26
	(100,100)	0.27	0.29	0.27	0.28	0.26	0.25	0.26	0.26
	(200,200)	0.33	0.36	0.33	0.34	0.29	0.29	0.29	0.29
	(250,250)	0.35	0.36	0.35	0.35	0.25	0.25	0.25	0.25

some misclassification. The performance on the augmented dataset was lower, with accuracy consistently at 0.34 to 0.35.This model struggled more with augmented data but maintained relatively stable performance across different resolutions. The Decision Tree model performed the lowest, with accuracy from 0.27 to 0.35 on the actual dataset and lower on the augmented dataset.The Random Forest model was the most consistent and reliable among the models tested, maintaining high performance across different resolutions.

After analyzing the results, it was found that higher resolutions of images generally improved ability of models to capture fine features and consequently enhancing classification accuracy. However, enhanced resolutions also incurred higher computational costs and processing time.

6.4 Evaluation

Based on the above results of all models, the CNN model was found to be the best. This model is further tested on the test dataset and it achieved a total 80% accuracy rate representing high stability on recognition of skin lesion images. The Figure 13 confusion matrix for test results corroborates the good performance of the Proposed CNN model with 96 correct predictions out of 118 total.

The performance of the model was considerably varying in different classes, as shown in the classification report. Actinic keratosis and seborrheic keratosis were difficult to classify as their precision, recall, F1-score values were 0.00. The vascular lesion class exhibited the highest performance.

6.4.1 Performance evaluation with existing studies

Comparing our results with established methodologies described in the literature is crucial for assessing the performance of our models. The work of Esteva et al. (2017), who used a CNN model to categorize skin lesions with excellent accuracy and beyond dermatologist level expectations, is one of the greatest recent instances of such a method. The CNN model fared remarkably well in this study as well, achieving an overall accuracy of 80%, which has been reported to be in line with earlier CNN-based studies. While our GNN

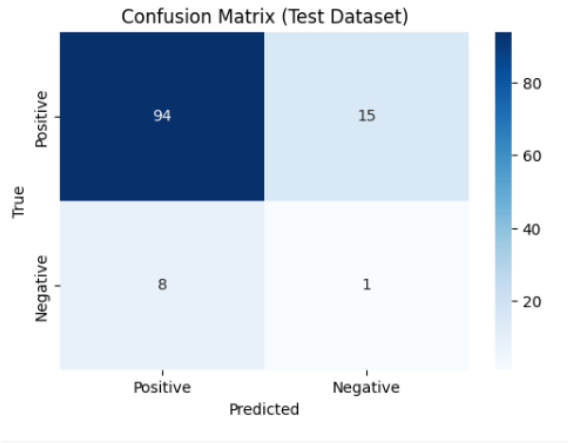

Figure 13: Test data Confusion matrix for CNN model

model only achieved 48% accuracy, which was not as excellent as the previously stated CNN, it is still useful for jobs involving transformer-based models, as described by Xin et al. (2022). Furthermore, the study of (AL-SAEDİ and Savaş; 2022) have shown that DenseNet and MobileNetV2 could achieve an accuracy level of up to 99.6% for balanced datasets. This comparison shows the benefits of utilizing CNN for the classification of skin cancer while also emphasizing key areas that still require improvement when using GNN or traditional classifiers. This suggests that while GNN would require improvement to reach equivalent performance levels, CNN might be a great tool for consistently handling this problem.

6.5 Discussion

The experiments in this study demonstrate what various machine learning models for skin cancer classification can or cannot do. The overall accuracy of the CNN model was 80% on the test data, but there were significant differences in terms of performance between different classes: common lesions like basal cell carcinoma were extremely well classified while rarer types such as actinic keratosis and seborrheic keratosis proved to be difficult.The study by (Esteva et al.; 2017) also indicated the prevalence of the most common classes, which is an inherent problem with uneven datasets.

The results demonstrate that the GNN model did not perform as well as the CNN due to the distortions introduced through conversion of images into graph representations using superpixels.The findings by (Scarselli et al.; 2008) support the need for better preprocessing techniques and use of more advanced GNN architectures like (GCN) Graph Convolutional Network and Graph Attention Network(GAN).The traditional machine learning models like Logistic Regression and Random Forest did not perform as well as CNN, which is consistent with other studies such as (Medhat et al.; 2022) showing that deep learning models are better at handling complex image data .However, these models were unable to identify tiny patterns, indicating a fundamental flaw in high dimensional picture processing.

Several modifications have been proposed to enhance the design. To deal with class imbalance and improve model generalization, SMOTE or GAN can be used as advanced data augmentation techniques. The inclusion of transfer learning from pre-trained models

on larger datasets could facilitate better feature extraction, especially for rare classes, as suggested by (Ashfaq and Ahmad; 2023). Furthermore, deeper and more complicated GNN architectures may be recommended that leverage the strengths of both methods by combining them with CNN for hybrid models.The methods of batch normalization and dropout regularization need to be adjusted to prevent overfitting, which is evident in the inconsistent validation accuracy seen in the study.In conclusion, while skin cancer classification gives promising results for CNN, these models must address class imbalance and increase their robustness using advanced techniques.

7 Conclusion and Future Work

This research successfully explored the potential of GNN for skin cancer image segmentation and classification and also comparing their performance with CNN and traditional machine learning models. The superiority of CNN can be seen in our results because they scored the highest overall accuracy at 80%, thus making them a good tool for categorizing skin lesions. On the other hand, GNN were unable to convert images into graphs, which restricted them to 48% only.Different methods such as Random Forest and Logistic Regression also gave satisfactory outcomes but were beaten by CNN in handling complex image data.In this research, the importance of alleviating class imbalance and improving model generalization is emphasized.For this reason, we had to consider data augmentation, which helped us balance our dataset although it was sophisticated and had negative effects on how the models work.

In future the research should focus on several key areas to further advance the field. One of them is exploring more sophisticated data augmentation techniques such as Generative Adversarial Networks(GAN) or (SMOTE)Synthetic Minority Oversampling Technique, which could perform better in dealing with class imbalance. In addition, integrating transfer learning using pretrained models on larger and more diverse datasets can lead to better feature extraction especially for rare classes.Further refinement of GNN is necessary, possibly through the development of hybrid models that combine the strengths of both CNN and GNN. Another potential improvement can come from adding multimodal data, combining clinical information with images to improve diagnostic capabilities. Also these models have great potential in terms of commercialization that are userfriendly diagnostic tools for dermatologists as they enable reliable and efficient early skin cancer detection. This research provides a strong ground for future developments in applying deep learning in medical image analysis with an ultimate aim to improve patient outcomes during skin cancer treatment.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Süsstrunk, S. (2010). Slic superpixels, *ResearchGate [Preprint]* .
URL: *https://www.researchgate.net/publication/44234783$_S$LIC$_s$uperpixels*

AL-SAEDİ, D. and Savaş, S. (2022). Classification of skin cancer with deep transfer learning method, *Journal of Biomedical Research* **15**(1): 202–210.

Aldwgeri, A. and Abubacker, N. (2019). Ensemble of deep convolutional neural network

for skin lesion classification in dermoscopy images, *International visual informatics conference*, pp. 214–226.

Aljohani, K. and Turki, T. (2022). Automatic classification of melanoma skin cancer with deep convolutional neural networks, *AI* **3**(2): 512–525.

Almaraz-Damian, J.-A., Ponomaryov, V., Sadovnychiy, S. and Castillejos-Fernandez, H. (2020). Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures, *Entropy* **22**(4): 484.

Ameri, A. (2020). A deep learning approach to skin cancer detection in dermoscopy images, *Journal of Biomedical Physics and Engineering* **10**(6): 801.

Ashfaq, M. and Ahmad, A. (2023). Skin cancer classification with convolutional deep neural networks and vision transformers using transfer learning, *Proceedings of the 2023 International Conference on Computer Vision and Pattern Recognition*.

Cullell-Dalmau, M. et al. (2021). Convolutional neural network for skin lesion classification: Understanding the fundamentals through hands-on learning, *Frontiers in Medicine* **8**.
URL: *https://doi.org/10.3389/fmed.2021.644327*

Dinesh, P., Vickram, A. S. and Kalyanasundaram, P. (2024). Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: Svm, knn, logistic regression, random forest and decision tree to measure accuracy, *AIP Conference Proceedings*.
URL: *https://doi.org/10.1063/5.0203746*

Elgamal, M. (2013). Automatic skin cancer images classification, *International Journal of Advanced Computer Science and Applications* **4**(3).

Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H. and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks, *Nature* **542**(7639): 115–118.

Hosny, K., Kassem, M. and Foaud, M. (2018). Skin cancer classification using deep learning and transfer learning, *2018 9th Cairo international biomedical engineering conference (CIBEC)*.

IEEE (2018). Improved adam optimizer for deep neural networks, *IEEE Conference Publication*.

Medhat, S., Abdel-Galil, H., Aboutabl, A. and Saleh, H. (2022). Skin cancer diagnosis using convolutional neural networks for smartphone images: A comparative study, *Journal of Radiation Research and Applied Sciences* **15**(1): 262–267.

Okuboyejo, D., Olugbara, O. and Odunaike, S. (2013). Automating skin disease diagnosis using image classification, *Proceedings of the world congress on engineering and computer science*, Vol. 2.

Raju, V. et al. (2020). Study the influence of normalization/transformation process on the accuracy of supervised classification, *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*.

Reshi, A. et al. (2021). An efficient cnn model for covid-19 disease detection based on x-ray image classification, *Complexity* .

Rustam, F. et al. (2022). Vector mosquito image classification using novel rifs feature selection and machine learning models for disease epidemiology, *Saudi Journal of Biological Sciences* **29**(1): 583–594.

Scarselli, F. et al. (2008). The graph neural network model, *IEEE Transactions on Neural Networks* **20**(1): 61–80.

SkinCancerISIC (n.d.).

Soudani, A. and Barhoumi, W. (2019). An image-based segmentation recommender using crowdsourcing and transfer learning for skin lesion extraction, *Expert Systems with Applications* **118**: 400–410.

Wu, L. et al. (2022). Graph neural networks: foundation, frontiers and applications, *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
URL: *https://doi.org/10.1145/3534678.3542609*

Xin, C., Liu, Z., Zhao, K., Miao, L., Ma, Y., Zhu, X., Zhou, Q., Wang, S., Li, L., Yang, F., Xu, S. and Chen, H. (2022). An improved transformer network for skin cancer classification, *Computers in Biology and Medicine* **149**: 105939.

Zia Ur Rehman, M., Ahmed, F., Alsuhibany, S., Jamal, S., Zulfiqar Ali, M. and Ahmad, J. (2022). Classification of skin cancer lesions using explainable deep learning, *Sensors* **22**: 6915.