National College of Ireland

Skin Lesion Classification Based on Various Machine Learning Models Explained by Explainable Artificial Intelligence

MSc Research Project
MSc. Data Analytics

Sarthak Gupta
Student ID: x20247575

School of Computing
National College of Ireland

Supervisor: Mr. Aaloka Anant

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sarthak Gupta

Student ID: x20247575

Programme: MSc. Data Analytics **Year:** 2022-2023

Module: Research Project

Supervisor: Mr Aaloka Anant

Submission Due Date: 01.02.2023

Project Title: Skin Lesion Classification Based on Various Machine Learning Models Explained by Explainable Artificial Intelligence

Word Count: 7656 **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: Sarthak Gupta

Date: 31.01.2023

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 Lesion Classification Based on Various Machine Learning Models Explained by Explainable Artificial Intelligence

Sarthak Gupta
x20247575

Abstract

Over the past few years, many lives have been lost due to the spread of various deadly diseases. There has also been a big rise in deaths from skin cancer, and most of these people could have lived longer if the cancer had been detected at an early stage. Therefore, a skin lesion classification model is proposed in this project, which will help in classification of skin lesions. A major problem with using artificial intelligence in the medical sector is a lack of transparency. Therefore, Explainable Artificial Intelligence (XAI) is used to explain the proposed classification model built using deep learning and machine learning models such as Convolutional Neural Networks (CNN) and Extreme Gradient Boosting (XGB Classifier). Two explainable artificial methods, such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive Explanations (SHAP), are used to explain the proposed models. The Convolutional Neural Networks (CNN) model performed on the augmented images produced an accuracy of 75.94%, which outperformed the other two models performed in this research. The explainable artificial intelligence method (SHAP) was used to explain the attribute importance, and the Local Interpretable Model-Agnostic Explanation (LIME) was used to explain how the model interprets images to predict its class.

***Keywords**: Explainable Artificial Intelligence (XAI), Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive Explanations (SHAP)*

1 Introduction

1.1 Motivation and Background

In recent years, there has been an increase in the number of deaths reported as a result of skin cancer. The fact that patients are not receiving the appropriate treatment at the appropriate time is the primary cause of this problem. The vast majority of people neglect it by treating it as a skin rash, which, if it is not diagnosed and treated at the right time, can develop into a skin lesion. Melanoma is a form of skin cancer that originates in the skin, while nonmelanoma skin cancer originates from the surface of the skin. Although there are a wide variety of subtypes of skin cancer, they are typically separated into these two categories. It is of the utmost importance to identify the kind of skin lesion present. As a result, for the purpose of this project, a classification model has been proposed that will help in the classification of skin lesions. The classification model can help reduce the time taken at the testing centres to identify the type of skin lesion. However, obtaining approvals is extremely

difficult as there is not much transparency between the machine learning models and how they make their decisions. This is also known as the "black box."

Explainable Artificial Intelligence, also known as XAI, is a method in machine learning that tries to decode the black box. The goal of explainable artificial intelligence is to explain a machine-learning model for humans so we can understand how a machine-learning model makes its decisions. As a result, explainable artificial intelligence will be applied to the proposed classification model to classify skin lesions, which will help in understanding how the models make classification decisions. This can also help medical practitioners understand the skin lesion better.

1.2 Research Question

How accurately can skin lesions be classified, and can the decision-making process of machine learning models be explained using explainable artificial intelligence?

1.3 Research Objective

The objective of this project is to build a classification model that classifies skin lesions accurately and efficiently. As a result, various deep learning and machine learning methods are proposed in this project. The research also aims to explain the models proposed in this project using explainable artificial intelligence methods. The following will be done to achieve the objective of the project: pre-processing of the data followed by transformation, where the dataset is split for testing and training the proposed CNN models and XGB Classifier model. comparison of models performed; Explanation of models proposed using explainable artificial methods.

1.4 Document Structure

This document is presented in various sections and subsections. The introduction section presents information on the research question and the motivation behind the research. The literature review section presents lessons from journals related to the proposed work. The research methodology section shows how the KDD steps are followed in this project. The section on Design Specification describes the architectures of proposed models and explainable AI methods. The Implementation section shows how the project is implemented. Evaluation section show the model evaluations and model interpretations (XAI).

2 Related Work

In this section, the supporting literature for the proposed project is examined: First, the literature on how conventional machine learning algorithms is used for image classification are reviewed, followed by the literature on how model interpretation is beneficial for the medical industry. followed by reviews of publications attempting image classification using explainable artificial intelligence.

2.1 Traditional methods used for image classification

A classification of skin cancer was attempted by *(Dubal et al., 2017)*. In this paper, the authors proposed a model to detect and classify skin cancer into six different types. The authors' dataset consisted of 463 images that were pre-processed before use. To extract the features from the images, the authors used the Asymmetry, Border, Colour and Diameter (ABCD) rule. The authors then applied a neural network model to the dataset, which was separated into training, testing, and validation phases (80, 10, and 10 percent, respectively). The dataset used in this project is very small, and the neural network could have been trained with more hidden layers to improve accuracy; however, the accuracy of the model trained by the authors was 76 percent.

The authors *(Daghrir et al., 2020)* of this journal are focusing their attention on a form of skin cancer known as melanoma. The authors of this project proposed a hybrid model for the detection of cancer that was based on semi-supervised learning. The authors used the well-known public dataset ISIC in this project and divided it into a training set and a testing set. The authors then fitted a hybrid model to detect cancer. The authors developed three different models: a machine learning model called KNN, a support vector machine model called SVM, and a deep learning model called CNN. The predictions of these models are combined using majority voting in the system that has been proposed. As the authors model worked by combining the prediction of the three models' prediction accuracy of models could have been increased before combing the model's output as the KNN model just produced an accuracy of 57 percent.

In this project, the authors *(Hosny et al., 2018)* have proposed a deep learning-based skin classification model. The authors trained and validated their model using the PH2 data set. The authors have also performed data augmentation and fine-tuning on the image dataset before proposing a transfer learning model. The authors also resized the images so that the model can be executed without difficulty. AlexNet was used to classify the images into three distinct skin cancer types: atypical nevi, common nevi, and melanoma. The results of the AlexNet model were 98.33 percent, which was significantly higher than the traditional model. The AlexNet model was validated using precision, sensitivity, and specificity by the authors.

As melanoma has become the most common form of skin cancer, for the detection of melanoma, the authors *(Jagadish Kumar et al., 2021)* of this journal proposed a machine learning model and a deep learning model in two stages to detect skin cancer. The image dataset used was resized to 512*512; these images were then augmented, where they were randomly cropped, colour adjusted, and flipped. To run the models, XGB Classifier and EfficientNet, the dataset was divided into training and testing segments. The models were tested on 10981 images, which produced an accuracy of 85 percent.

The authors *(Li et al., 2014)* created an image classification model. The classification model proposed by the authors is built to classify patches in the lungs. The authors used a relatively

small dataset of just 2062 images to build and test their deep learning models. The authors proposed a CNN model. The model was inputted with the original dataset. This model proposed by the authors performed the best when compared to other models.

This research article *(Dildar et al., 2021)* proposes an image classification model that is used to classify skin cancer. The authors' extracted images contain information on skin infections. These extracted images are then processed and split into a testing set and a training set. The authors then proposed four deep learning and machine learning models such as CNN, ANN, KNN, and GAN. The models were first trained on 3797 images and then tested to make predictions. The KNN model produced an accuracy of 71.20%, the ANN model was 63% accurate, the GAN model produced 70% accuracy, and the CNN model performed at 86 percent accuracy. The CNN model performed by the authors in this project outperformed the rest.

2.2 Need for model interpretation

Model explainability can be extremely beneficial in the medical field, as it allows researchers to explain their results and predictions. As discussed in the authors' article, "Using artificial intelligence to advance the field of medicine," explanations of machine learning model results are required. The authors *(Tjoa and Guan, 2021)* discussed how explainable artificial intelligence (XAI) can help machine learning models fill in missing details. In addition, they discussed the various types of interpretabilities that can be used to describe machine learning and deep learning models. No experiment was carried out in the research articles. However, the authors mention which explainable artificial intelligence library can be used for a particular model.

The difficulties that artificial intelligence encounters in the field of medical science were published by the authors *(Kelly et al., 2019)* of another journal. According to what has been discussed in this journal, the most significant barrier that stands in the way of artificial intelligence achieving success in the field of medicine is the interpretation of models. The authors place a strong emphasis on the fact that machine learning models should not be biased toward a single feature or attribute, and consequently, the model should be human interpretable.

2.3 Explainable Artificial Intelligence

To understand the machine learning model and how it makes predictions, this is also known as the "black box." The authors *(Bhandari et al., 2022)* of the journal have used explainable artificial intelligence to interpret the model's predictions. In this project, the authors have built a classification model that classifies pulmonary disorders using deep learning models. The authors then used explainable AI libraries such as LIME, SHAP, and Grad CAM to interpret the deep learning models. The experts were successfully able to classify the images

into those of tuberculosis, pneumonia, and COVID-19, and the CNN model produced an accuracy of 94 percent. They used SHAP, LIME, and Grad CAM to understand the classification process of the CNN model.

In an attempt to explain deep learning algorithms built to detect and classify retinal disease, the authors *(Reza et al., 2021)* of this paper first propose a deep learning model that predicts and classifies retinal disease using deep learning models. The dataset used by the authors in this project is fairly large. The image is first preprocessed and augmented, and then three deep learning models are applied: inseptionv3, resnet50, and efficient net. The predictions of these models are then fed into the LIME to explain the results. In this project, the authors focused on LIME to explain the misclassification of retinal diseases.

To understand the "black box" of deep learning and machine learning models, the authors *(Shakil and Rabiul Alam, 2022, p.)* in this journal attempted to explain the model results using the explainable library SHAP. In this project, the authors proposed a classification model based on a hybrid of CNN-LSTM and machine learning models such as the Random Forest and Extra Tree algorithms to classify toxic voices. Text data is used by the authors to build the classification model. The authors then used the SHAP library to explain the model, as SHAP can be seen to interpret feature contribution in a machine-learning model. The authors created plots to demonstrate the impact of features on model prediction.

In this paper, the authors *(Gezici and Tarhan, 2022)* propose a software defect detection model built using a gradient-boosting classifier. The authors of this paper focused more on the explainability of the model. They used three different, explainable artificial methods to explain the results. The methods used by the authors are LIME, SHAP, and ELI5 (explain like I am a 5-year-old) to interpret the gradient boost model. The model performed in this project produced an accuracy of 84 percent, and the model's results were fed to the three model explainers. These three model explainers then produced reasons and explanations for the prediction made by the model.

The above journals are reviewed to understand how image classification using various deep learning and machine learning techniques can be used to build a classification model to classify skin lesions. It was also discovered how image augmentation helps build a better classification model. This was followed by reviews of papers explaining the need for model explainability in the medical sector. Some attempts by other authors were made to explain their machine learning model. LIME and SHAP were the most common methods used by the authors to explain their models.

3 Research Methodology

The Knowledge Discovery in Databases (KDD) process is followed in this project. The KDD steps followed in this project are data collection, data pre-processing, data transformation, data mining, and evaluation. To understand the machine learning and the deep learning model, an additional step called "Model Explanation" is added after the KDD steps mentioned.

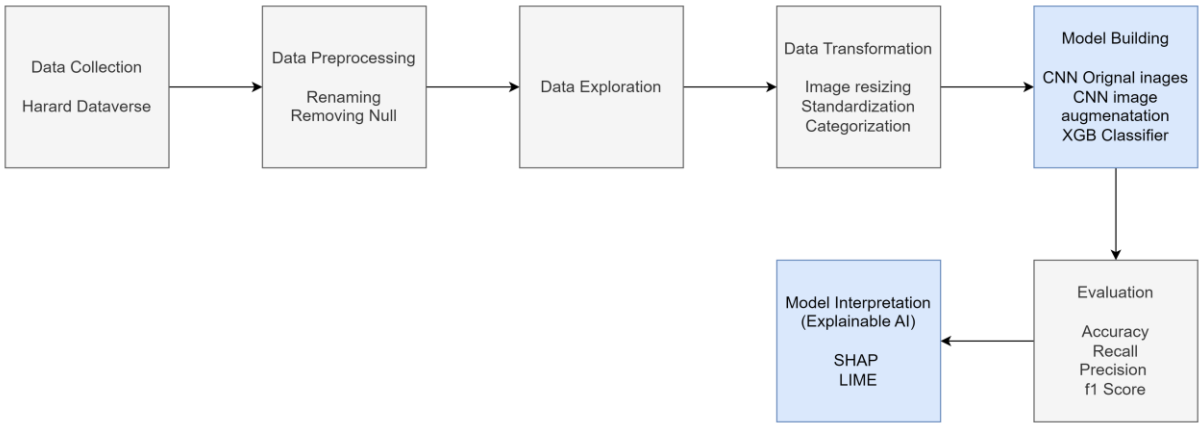

Figure 1: Architecture Diagram

3.1 Data Collection

The initial step is to collect data. Many datasets are available over the internet; however, for this project, the dataset was downloaded from the Harvard Dataverse website. The dataset is accessible to the general public and contains a vast collection of dermatoscopic images of frequently encountered skin lesions. This dataset contains a total of 10015 images of various skin lesions, which are divided between two folders. This dataset has a sufficient number of images to test and train a good classification model. The lesion images in this dataset were collected from people all over the world.

The CVS file contains seven attributes, such as lesion id, image id, dx, dx type, age, sex, and localization. Some column names are changed for better understanding.
Lesion ID: Multiple images of a specific lesion are provided. The lesion ID can be used to reference these images.
Image ID: Each image in the dataset possesses a unique image ID. This attribute is subsequently used to map images to their corresponding image ids.
lesion_type: This column provides details regarding the type of lesion. The dataset contains seven distinct types of skin lesions, with an aberration assigned to each type.
lesion_confirmation_type: The column contains confirmation information for the specific lesion. Histopathology, follow-up examination, expert consensus, and confocal microscopy are among the four categories.
Age: age of the person.
Sex: Gender of the person.
lesion_location: This column contains information about where the lesion was detected on a person's body.

3.2 Data Pre-processing

Following data collection, the dataset must be pre-processed so that it is ready for further use. In this project, the following steps are taken to prepare the dataset.

The initial CSV file just contains information on the skin lesion. This dataset is then transformed, and some new columns are added to help understand the dataset better. As the images in the dataset were divided into two separate folders, the two folders were merged. A new column is created to map the path of the image to the correlating image id. Another column is created to store the full name of the type of skin lesion, as the initial column only shows the abbreviation of each skin lesion type. These skin lesion types are then converted to categorical values and saved in a new column. The dataset is then checked for null values, and the column "age" was found to have 57 missing values. These missing values were then replaced with the column age mean.

3.3 Data Exploration

To gain a deeper understanding of the information contained in the dataset. It is analysed using methods of data visualization. The key findings from the dataset are represented below.

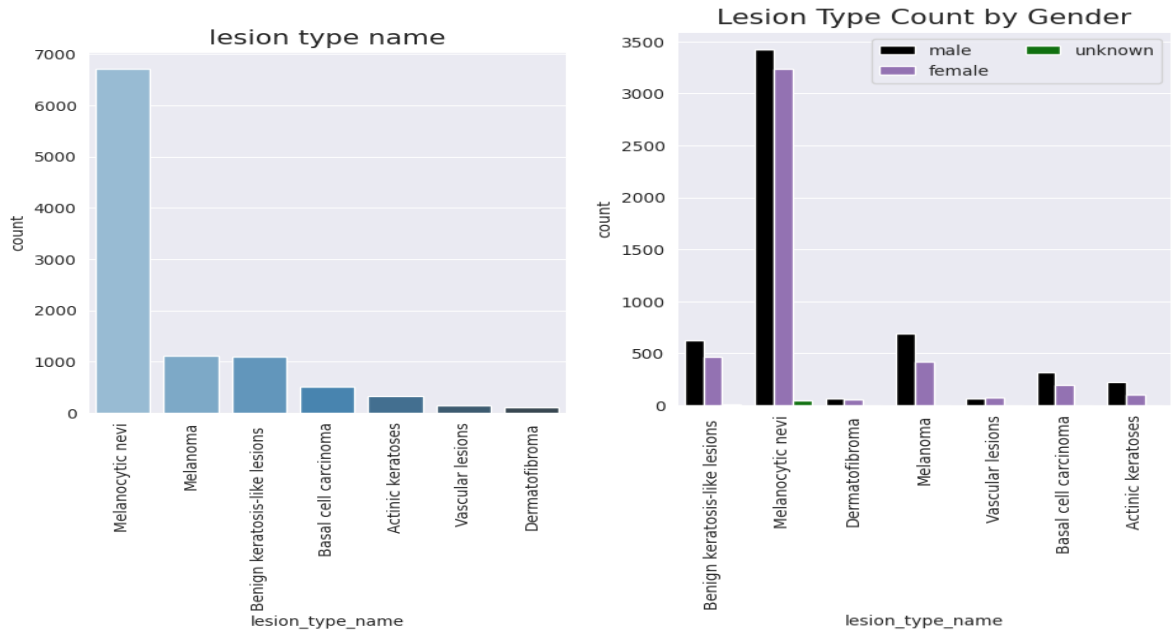

Figure 2: Lesion Type count

The above bar chart (Figure 2) was designed to illustrate the distribution of all seven types of skin lesions present in the dataset. It is evident from the preceding bar graph that the dataset contains a large number of skin lesions (melanocytic nevi) relative to the other variables. Melanoma and benign keratosis-like lesions have comparable numbers of images; similarly, actinic keratoses and basal cell carcinoma have comparable numbers of images. When the type of skin lesion is evaluated based on gender, comparable results are observed.

The bar graphs below (Figure 3) are created to analyze the age distribution of people affected by skin lesions. From the bar graph created, it can be seen that people between the ages of 35 and 70 are affected by some kind of skin lesion. However, people between the ages of 45 and 50 are most affected by some types of skin lesions. When the same graph is used with a gender distribution filter, a similar trend is observed. However, more women between the ages of 45 and 50 are affected than men of comparable age.

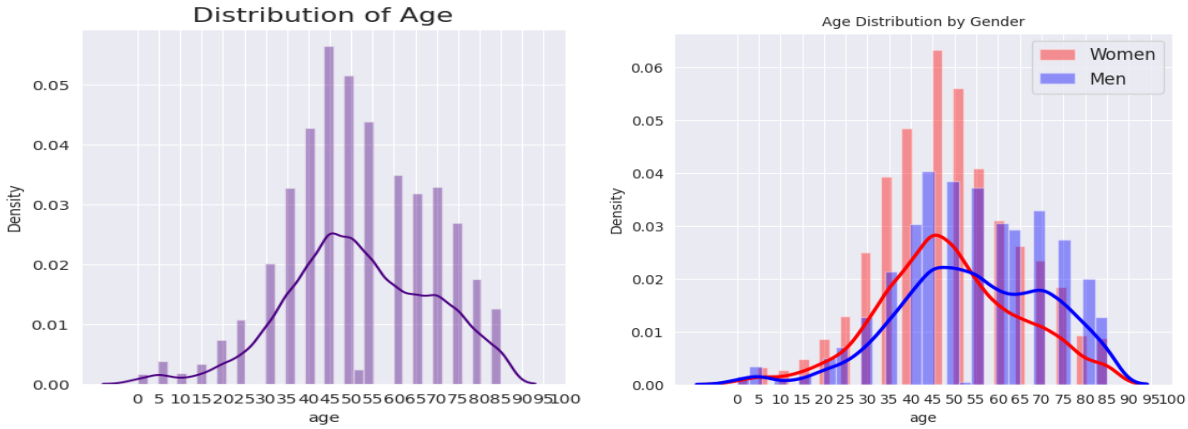

Figure 3: Age Distrubition Curves

The bar graph blelow (Figure 4) shows the distribution of skin lesions found on each body part. The graph clearly states that most skin lesions are found on the back and lower extremities of the body. A significant amount of skin lesions are found on the trunk, upper extremities, and abdomen of the body. However, a lower number of skin lesions are found on Acral i.e., body parts such as the hands, arms, foot, ears, and nose.

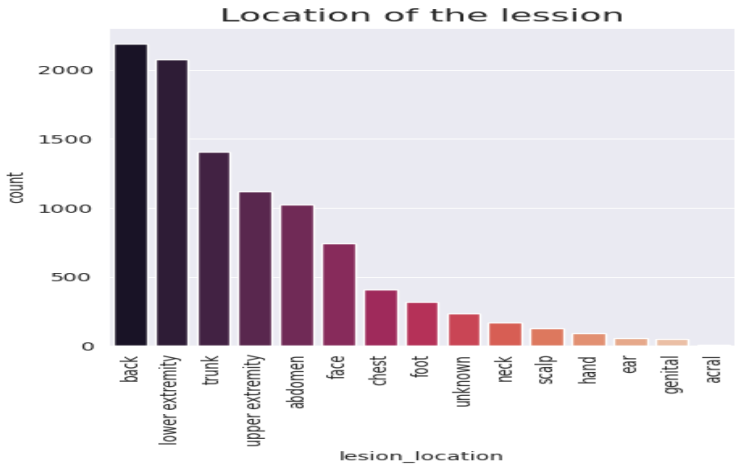

Figure 4: Lesion location Graph

The images below (Figure 5) represent the skin lesions belonging to a particular type of skin lesion. Four images from all seven types of skin lesions are represented in the image above.

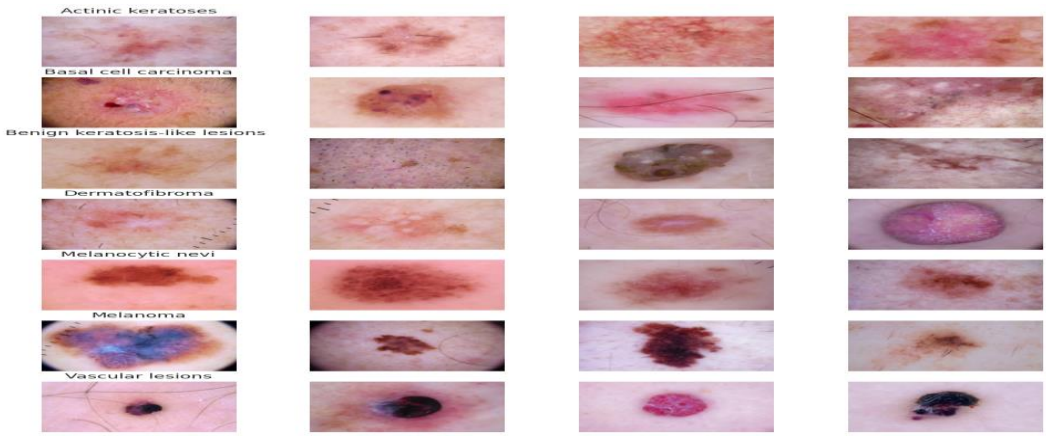

Figure 5: Lesion images

3.4 Data Transformation

The pre-processed dataset is then transformed. This step involves preparing the dataset so that the machine learning and deep learning models proposed can be built using the transformed dataset. The images are resized. The dataset is then divided into two parts: test and training. followed by the standardization of image pixels and the categorization of the attributes. This transformed dataset is then used to build models and classify skin lesion types. This section is discussed in detail in the Implementation section.

3.5 Model Building

In this section, the transformed data is split into train and test in the ratio of 80:20. and the chosen models are carried out: the CNN model on the original dataset, another CNN model after image augmentation, and the XGB Classifier model. The specification and parameterization of these models are discussed in the next two sections, respectively. The model results are then evaluated.

3.6 Evaluation

In this section, the proposed models are evaluated. To evaluate the proposed model, the following evaluation matrix is used: Accuracy, Precision, Recall, and F1-Score. These are calculated using their respected formulas.

3.7 Explainable AI

In this section, the two approaches to explainable artificial intelligence, known as SHapley Additive Explanation (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), are both implemented. SHAP is used to understand the feature importance of each attribute in model prediction. The second method, LIME, is used to understand how the machine learning model interprets images to make predictions.

4 Design Specification

This section discusses the architecture of the proposed models. Two Convolution Neural Network (CNN) models are proposed in this project, the first of which is built on an original dataset and the second of which is built on images with augmented data and adjusted parameters. followed by the third model proposed, the Extreme Gradient Boost classifier. After model building and evaluation, the proposed model is explained using the explainable artificial intelligence libraries SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), which are discussed in this section.

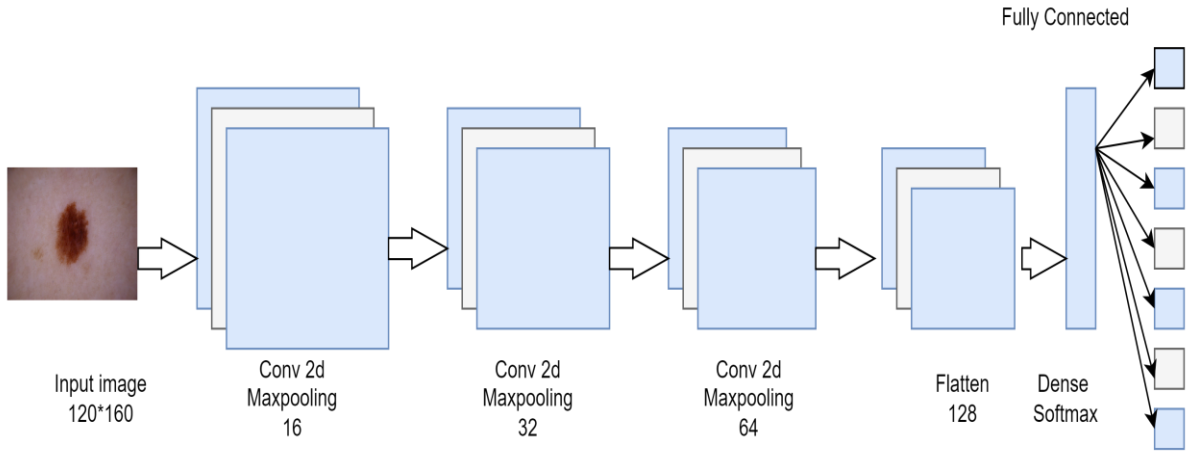

Figure 6: CNN Architecture

4.1 Convolutional Neural Network (CNN)

A convolutional neural network is a deep learning model that learns directly from the data. Convolutional neural network models are the most used for image class prediction. A model can be built with many CNN layers. The CNN layers consist of a convolutional layer followed by pooling layers and fully connected layers. The proposed model is implemented sequentially; the layers are added one by one and then compiled. Each layer consists of an important parameter called the activation function, and depending on the type of prediction, a different activation function is used. The layers used in the proposed CNN model are described below.

Convolution Layer: This is the first layer that is used to extract features from the input layer provided to the model. This layer must also define the weight and type of activation function used as This layer may also serve as input for the next layer. Three convolution layers are used.

Batch normalization: This layer is added after the activation layer to standardize the output and maintain data distribution. Four layers are added to the CNN model proposed.

Max Pooling Layer: This layer is the bridge between the convolution layer and the fully connected layer. This layer takes the largest element from the feature and generalizes it. Three Max pooling layers are added to the proposed CNN model.

Dropout layer: This layer is added to overcome this issue of overfitting in the model. Therefore, four dropout layers are added to the proposed CNN model.
Flatten Layer: This layer is used for converting the data from the previous layers to be passed to the fully connected layer. As a result, the CNN model gains a flattening layer.
Dense layer: This layer is used for classification of images based on outputs from the previous layers. Two dense layers are fitted to the proposed CNN model.
Activation Function ReLU: The activation parameter determines whether a neuron should be activated or not. ReLU removes all the vector values and replaces them with zero, which controls the computing power required to operate the model. All the hidden layers in the proposed model use ReLU as an activation function.
Activation Function SoftMax: This activation function is used in the fully connected layer. A SoftMax activation function is used to create a multiclass classification model. Therefore, the SoftMax activation function is used in the last layer of the proposed CNN model.

4.2 Extreme Gradient Boost

This It is a machine learning algorithm which is based on gradient tree boosting algorithm. This algorithm has built it capabilities which handles null values. This model assists in lowering the error rate, which in turn assists in preventing the model from becoming overfit. In this project XGBoost Classifier is used to build a classification model and predict type of skin lesion. The training set is used to fit the model, and then the test set is used to validate the model's predictions.

Following the model's building and evaluation, this project aims to understand the model's "black box." In this part of the project, the decision-making process of the machine learning model will be interpreted. To understand the model's decision-making process, explainable artificial intelligence (XAI) methods are used, such as SHapley Additive exPlanation (SHAP) and Local Interpretable Model-agnostic Explanations (LIME).

4.3 SHapley Additive Explanation (SHAP)

SHAP is a state-of-the-art explainable AI framework that is used to understand the machine-learning models was published on 2017. SHAP is used to explain sophisticated machine learning and deep learning models, where the models are fed with features as input and produce some predictions as output. To underline the importance of the feature, "SHAP values" are calculated.

SHAP Value is a concept based on Game Theory, where this value quantifies the contribution of all the features used to build the classification model. The generalized formula to calculate the SHAP value *(Mazzanti, 2021)* for a particular feature is given below. SHAP is employed in this project as one of the explainable AI methods to understand the features' importance, which is fed into the three models proposed in this project.

$$SHAP_{feature}(x) = \sum_{set:feature \in set} [|set| \times \binom{F}{|set|}]^{-1} [Predict_{set}(x) - Predict_{set \setminus feature}(x)]$$

4.4 Local Interpretable Model-Agnostic Explanations (LIME)

This is another explainable artificial intelligence method to explain the "black box" of machine learning models. This method can be used to interpret many different types of models built with text, images, and tabular datasets. Therefore, LIME is used to explain how the images are interpreted by the models to make decisions.

To understand an image, this method first divides the image into "Superpixel" using segmentation. Then it creates perturbation data around the section of the image that needs to be explained, **Perturbation** data is generating “n" variations of the original image by hiding parts of the original image. This step is followed by, the **Prediction** a class for each perturbed image using the proposed models performed in this project. LIME then calculated the distance between the original image and all the perturbated images. LIME then calculates the importance of each perturbated image using the cosine function; this is termed "**Weights**."

This is followed by the final step, where the LIME uses the **Perturbation, Prediction, and Weights** calculated in previous steps to fit a linear regression model. The output of the linear regression model presents the part of the image that is responsible for the model's classification of this image into a particular class.

5 Implementation

Environment Setup

The implementation of this research work required a system with higher computational power; therefore, the implementation was performed on the Amazon Web Services provided by the college, An EC2 instance was created to implement the project. The system was built with 75 GB of memory and a 4 Tesla V4 16 GB GPU.

5.1 Data Preparation and Transformation

The dataset is imported using the pandas library in Visual Studio. Column names were changed, and new columns are added to the dataset, such as "path", "lesion_type_name", and "lesion_type_categorical". The dataset contains 57 missing values in the age columns, which were replaced with the mean of the column age. The “lesion_type” column was then converted into a categorical column. Following this, the dataset was visualized with the help of libraries such as Seaborn and Matplotlib. The key findings from the dataset are discussed in the Research Methodology section.

The original size of the images in the dataset was 450*600*3. This was then resized to 120*160*3 (height, width, channels) as the original size of the image was too large to provide it as the input to the models and required more computational power to run the proposed models.

The dataset was divided into two parts: a testing set and a training set, in a ratio of 80 percent to 20 percent. This is followed by the standardization of the test set and the training set. First, the mean and standard deviation were calculated for both the test and the training set, then the mean was subtracted from the original and divided by the standard deviation to standardize the dataset. Using one hot encoding, the values of test and train are converted to categorical. The train set was further divided into an 87 percent train set and a 13 percent validation set.

The final size of the train set is 6970, the test size is 2003, and the validation size is 1042.

5.2 Convolutional Neural Network on Original Images

The proposed CNN model is built and tested on the transformed dataset. Four convolutional layers are added to the proposed CNN model, with filter sizes of 16, 32, 64, and 128 in the respective four layers. Three convolutional 2D layers are used to extract features from the input layer or the previous layer. Four batch normalization layers are added to standardize the output and maintain data distribution. Three maxpooling layers are added to connect the convolution layer with the fully connected layer by extracting the largest element of the image and generalizing it. A Flatten layer is also added to convert the output from the previous layer so that it can be passed to the fully connected layer. Four dropout layers are added to reduce overfitting. All the hidden layers are activated using the ReLU function. The final layer, Dense is added for the image classification and used softmax to as activation. The architecture diagram of the model proposed is shown in Design specification.

The model was run with 60, 70, and 100 epochs. However, when the graph for model accuracy vs. epochs was examined, the graph became constant after 50 epochs. Therefore, the CNN model ran for 50 epochs. The batch size is set to 5 for this model and 10 for the next CNN model, as the system requires more GPU memory to run batch size higher than these. The proposed CNN model is fitted to the training and test sets and validated using the validation set.

5.2.1 Results for CNN model

The CNN model, as performed and tested on the test set, produced an accuracy of 73.89 percent. The model also produced an F1 score of 70, a recall value of 73, and a precision score of 71. The classification report for all seven lesion types is generated and represented below. The classification report reflects the fact that the dataset contained more images of melanocytic nevi, as discussed in the pre-processing section. The weighted average of the precision, f1 score, and recall are mentioned above. The area under the curve calculated for this model is 68.62 percent.

Model Classification Report

	precision	recall	f1-score	support
Actinic keratoses	0.41	0.38	0.40	60
Basal cell carcinoma	0.45	0.46	0.46	97
Benign keratosis-like lesions	0.54	0.42	0.47	224
Dermatofibroma	0.67	0.07	0.13	27
Melanocytic nevi	0.79	0.96	0.87	1320
Melanoma	0.62	0.10	0.17	246
Vascular lesions	0.64	0.79	0.71	29
accuracy			0.74	2003
macro avg	0.59	0.46	0.46	2003
weighted avg	0.71	0.74	0.69	2003

Figure 7: Model 1 Classification Report

Figure 8: Model 1 Accuray and loss Graph

The above two graphs are generated to show how the accuracy of the model and model loss have changed with respect to the epochs. From the first graph, it can be seen that both the training and validation accuracy of the model are increasing as the number of epochs increases, and an accuracy of 73.89 is achieved at the 50th epoch. The second graph depicts how the model loss has decreased with increasing epoch count, with the final loss value being 0.72.

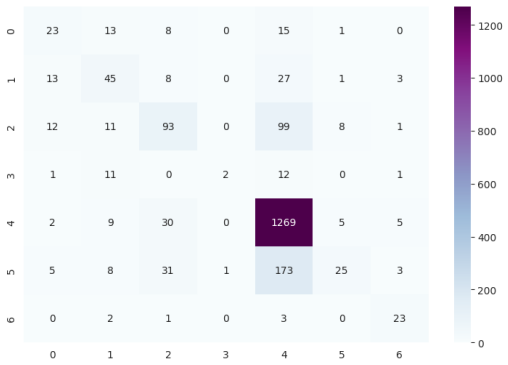

Figure 9: Model 1 Confusion Matrix

The confusion matrix is created to represent correctly predicted skin lesions from each class. There are seven categories of skin lesion and the above confusion matrix shows the number

of correctly predicted skin lesions for each class when the model was tested on the test set containing 2003 images.

5.3 Convolutional Neural Network After Image Augmentation

The second model proposed in this project is also based on the convolutional neural network architecture. However, this CNN model is applied to augmented images. This CNN model is implemented with the same four convolutional layers as discussed in the previous model. To implement this model, first augment the images in the training set with various parameters using Keras' ImageDatagenerator.i8

Image Augmentation: Various parameters were tried to achieve better results when augmenting the images. The image rotation range was set to 20, and a variety of rotations were tried; however, they made little difference. The width shift and the height shift of images are set to 0.1, which allows images to shift randomly right or left, up or down. The shift range value should be between 0 and 1. A value that does not disturb the content of the image should be provided as input. The zoom range of the image is set to 0.1; other values tried reduced the accuracy of the model. The images are also flipped horizontally. These modifications are fitted to the images of the training dataset.

ReduceLROnPlateau: This is a callback function applied to the model during the training process. The models usually benefit from reducing the learning rate once the learning becomes stable. This callback feature is then introduced. If the model has not improved since the "patience" parameter was set, The learning rate is reduced. For this model, the patience value is set to 3, i.e., if the model stagnates, the learning rate of the model is reduced.

This CNN model ran with 50 epochs, and the batch size was increased to 10 for this model as any value higher than this required more GPU memory to run the model. The model was fitted to an image-augmented training and testing set and validated on the validation set. This model was fitted with the ReduceLROnPlateau callback function to reduce the learning rate when model learning stagnates.

5.3.1 Results for CNN model After Image Augmentation

The accuracy of the second CNN model after image augmentation was 75.94 percent. In addition, the model obtained an F1 score of 72, recall value of 76, and precision score of 714. The classification report for each of the seven types of lesions is generated and displayed below. The calculated area under the curve for this model is 70.66. The CNN model performed more effectively than its predecessor.

Model 2 Classification Report

	precision	recall	f1-score	support
Actinic keratoses	0.46	0.37	0.41	60
Basal cell carcinoma	0.49	0.47	0.48	97
Benign keratosis-like lesions	0.60	0.46	0.52	224
Dermatofibroma	0.46	0.22	0.30	27
Melanocytic nevi	0.81	0.97	0.88	1320
Melanoma	0.71	0.19	0.30	246
Vascular lesions	0.63	0.76	0.69	29
accuracy			0.76	2003
macro avg	0.59	0.49	0.51	2003
weighted avg	0.74	0.76	0.72	2003

Figure 10: Model 2 Classification Report

The preceding two graphs illustrate the progression of the model's accuracy and model loss with respect to epochs. It can be observed from the first graph that both the training and validation accuracy of the model increase as the number of epochs increases, reaching 75.94 at the 50th epoch. The second graph illustrates how the model loss has decreased as the number of epochs has increased, with a final loss value of 0.62.

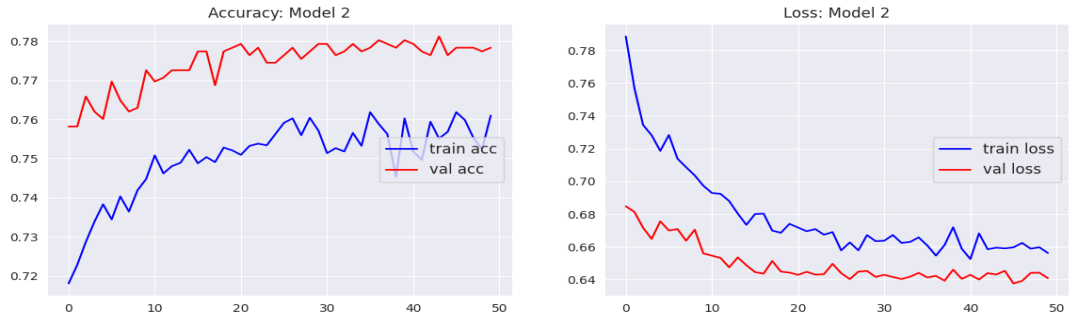

Figure 11: Model 2 Accuracy and Loss Plot

The confusion matrix is constructed to represent accurately predicted skin lesions within each class. There are seven categories of skin lesion, and the above confusion matrix displays the number of correctly predicted skin lesions for each class when the model was evaluated using 2003 images from the test set.

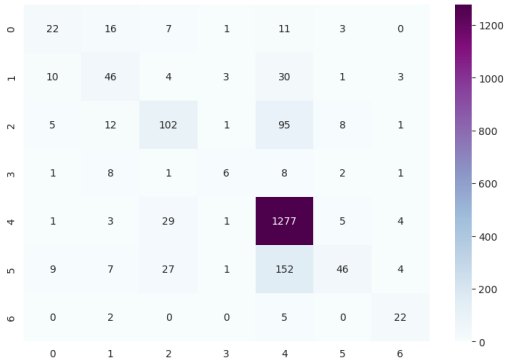

Figure 12: Confusion Matrix for Model 2

5.4 Extreme Gradient Boosting (XGB Classifier)

The third model proposed in this project is Extreme Gradient Boosting, also known as the "XG Boost." For this project, "XGB Classifier" is imported. First, a copy of the pre-processed dataset is used, and all the columns are converted into categorical values. then split the dataset into two parts: testing and training the target variable is "lesion type." The model is then fitted. and this model is tested on the test set.

5.4.1 Results for XGB Classifier

This model performed on the test and train dataset produced an accuracy of 71.92% when fitted to the test set. This model and prediction are used by SHAP to explain the feature importance and contribution in classification of skin lesions.

6 Evaluation

6.1 Model Evaluation

	CNN Original Dataset	CNN After Image Augmentation	XGBoost Classifier
Accuracy	73.89%	75.94%	71.92%

Table 1 Model Comparison.

The table above shows the comparison between the accuracy of all three models performed in this research project and the accuracy of the CNN model after image augmentation, which is 75.94 percent. which outperformed the other two models; the precision, recall, and f1 score of this model are 74, 76, and 72, respectively. The results of each model performed are discussed briefly in the Implementation section (5.2.1, 5.3.1, and 5.4.1). This is followed by the model explanation, which is another way to evaluate and understand the proposed models.

6.2 Model Interpretation (Explainable AI)

Model interpretation is an important part of this research, as this is done to evaluate the prediction models. Therefore, in this section, explainable artificial methods such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanation) are used to interpret the model performed in this project.

6.2.1 Shapley Additive Explanation (SHAP)

In this project, this explainable artificial intelligence framework is used to understand the feature importance of the attribute when developing a classification model to classify skin lesions. The SHAP library is first installed and imported into the environment. The Extreme Gradient Boost is given as the input to SHAP as the attributes in this model are precisely categorized which will help in understanding of the results produced by the SHAP explainer.

First, the SHAP value is calculated using the formula provided above. This value is then used to describe the importance of a feature when building the classification model.

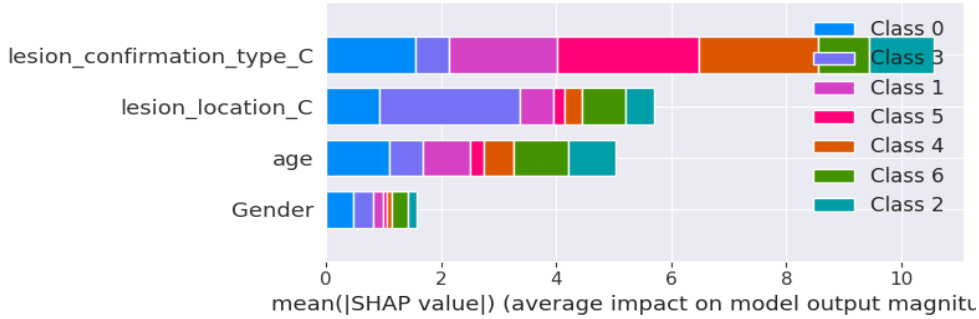

Figure 13: Feature Contribution grapgh

The graph shown above is a representation of each attribute used in the XGBoost classifier model to classify skin lesions. The graph shows the SHAP value corresponding to each attribute. The greater the SHAP value, the greater the contribution to the model. The graph also represents which attributes help in the prediction of which class. "lesion_confirmation_type" contributes most to the model. followed by the attributes "lesion location" and "age." "lesion_confirmation_type" contributed most to classifying lesions types 1, 5, and 4, whereas "lesion location" contributed most to classifying type 3.

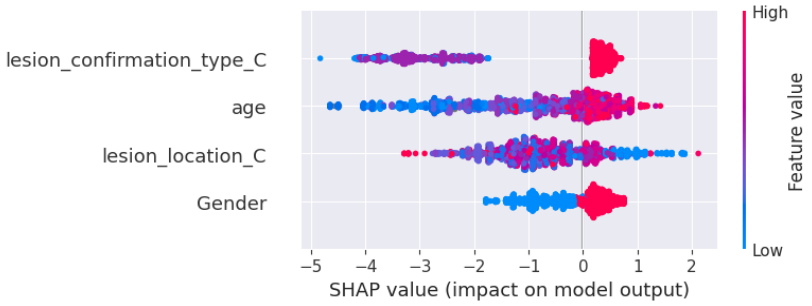

Figure 14: Summary Plot

From the summary plot created above, it can be seen what effect each attribute has on the SHAP value. It can also be seen that "lesion confirmation type", and "gender" are associated with a high SHAP value.

Figure 15: Feature Impact Graph

The above line graph shows which attributes' categories impacted positively and which impacted negatively on the model. As seen from the graph above, the attributes "skin confirmation type's" class 3 and "lesion location's" class 2 contribute positively to the model, whereas age 40 and gender 0 (female) have a negative impact when making predictions.

6.2.2 Local Interpretable Model-Agnostic Explanations (LIME)

This is the second method used to explain the decision-making process of the model. In this method, the CNN model is given as the input, and the focus is on explaining how the CNN model interprets images to build a classification model. First, the LIME library is installed and imported. The working of the LIME method to explain images is explained in the design specification section.

For this model, a random image is selected to explain how the CNN model interprets the image and predicts its class. The initial image is presented below, which if fed to the CNN model will predict its class. The following image is predicted to belong to Class 4.

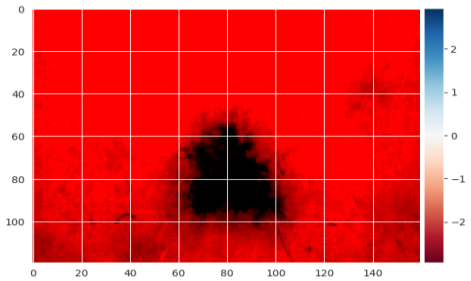

Figure 16: Orignal Image

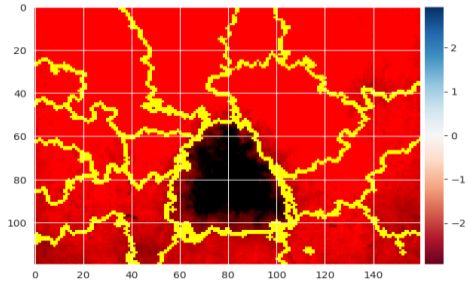

Figure 17: Segmented Image

The LIME method first generates a "Superpixel" from the image using segmentation. This image was divided into 16 sections, so there can be 16 possible **perturbations** of the original image, which is generating "n" variations of the original image by hiding parts of the original image. (n = 16) in this case. The following two images, represented below, are from the 16 possible **perturbations.**

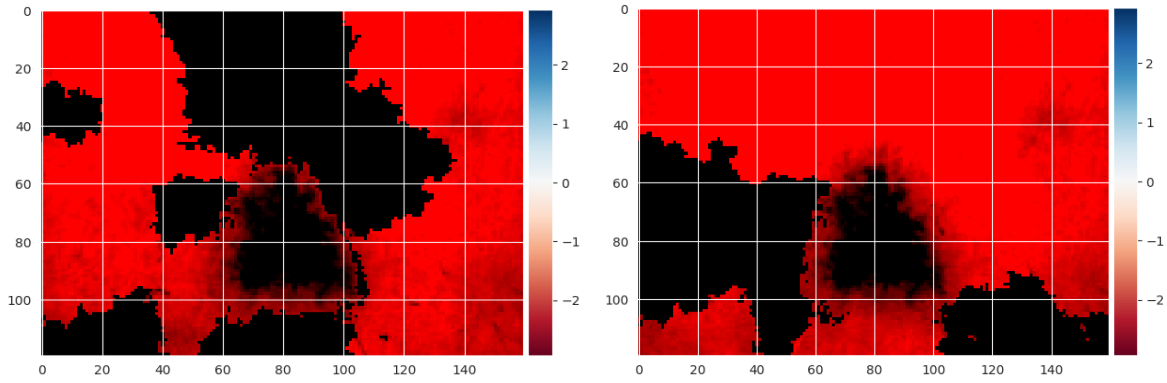

Figure 18: Perturbated Images

LIME then uses CNN to predict the class of each perturbation image. followed by calculating the distance of each perturbation from the original image. The importance of each perturbation image, also known as "weights," is also calculated using the cosine function.

The **Perturbation, Prediction, and Weights** calculated in previous steps are used to fit a linear regression model. The output of the linear regression model presents the part of the image that is responsible for the model's classification of this image into a particular class.

The following picture presents the segment of images responsible for classifying the image into class 4.

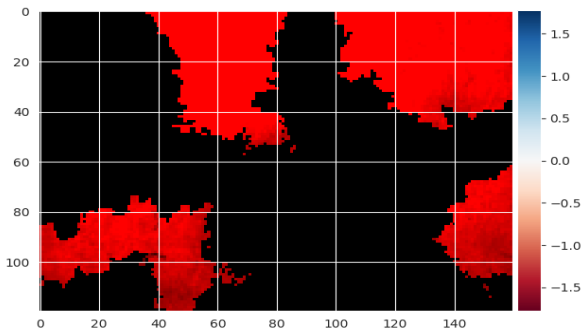

Figure 19: Section of image responsible for predicting the class

7 Conclusion and Future Work

To achieve the project's goals, a classification model for skin lesions is developed. The proposed models were built on a pre-processed and transformed data set. This study used three models: two CNN models and one other model. The first model was performed on the original dataset and had an accuracy of 73.89 percent. The second model built on the same CNN architecture was performed on an image-augmented data set; this model performed better and produced an accuracy of 75.94 percent. The third model of the XGB Classifier produced an accuracy of 71.92 percent. Out of the three models tested, CNN with augmented images performed the best with a precision, recall, and f1 score of 74, 76, and 72, respectively. The second part of the project was to explain the "black box" of these models used in this research. SHAP and LIME, two methods of explainable artificial intelligence, were used. where SHAP explained the importance of each feature when building a classification model, while LIME was used to explain how a model interprets images to predict its class. To conclude this project, the proposed research objectives were met. And to answer the research question, a skin lesion classification model can be built using machine learning and deep learning models, and explainable artificial intelligence helps to decode the black box of these models by explaining the decision-making process used to build the classification model.

The proposed models in this research were successfully able to classify skin lesions, and explainable artificial intelligence explained the decision-making process of the models. However, for the model to be able to be used in the medical sector, it needs to provide better results. The model explanations generated by explainable AI at times do not make complete sense to humans; therefore, this can be further explored to produce concrete explanations for the decision-making process. This will allow artificial intelligence to make a greater contribution to the medical sector.

References

Bhandari, M., Shahi, T.B., Siku, B., Neupane, A., 2022. Explanatory classification of CXR images into COVID-19, Pneumonia and Tuberculosis using deep learning and XAI. Comput. Biol. Med. 150, 106156. https://doi.org/10.1016/j.compbiomed.2022.106156

Daghrir, J., Tlig, L., Bouchouicha, M., Sayadi, M., 2020. Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach, in: 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). Presented at the 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp. 1–5. https://doi.org/10.1109/ATSIP49331.2020.9231544

Dildar, M., Akram, S., Irfan, M., Khan, H.U., Ramzan, M., Mahmood, A.R., Alsaiari, S.A., Saeed, A.H.M., Alraddadi, M.O., Mahnashi, M.H., 2021. Skin Cancer Detection: A Review Using Deep Learning Techniques. Int. J. Environ. Res. Public. Health 18, 5479. https://doi.org/10.3390/ijerph18105479

Dubal, P., Bhatt, S., Joglekar, C., Patil, S., 2017. Skin cancer detection and classification, in: 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI). Presented at the 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), pp. 1–6. https://doi.org/10.1109/ICEEI.2017.8312419

Gezici, B., Tarhan, A.K., 2022. Explainable AI for Software Defect Prediction with Gradient Boosting Classifier, in: 2022 7th International Conference on Computer Science and Engineering (UBMK). Presented at the 2022 7th International Conference on Computer Science and Engineering (UBMK), pp. 1–6. https://doi.org/10.1109/UBMK55850.2022.9919490

Hosny, K.M., Kassem, M.A., Foaud, M.M., 2018. Skin Cancer Classification using Deep Learning and Transfer Learning, in: 2018 9th Cairo International Biomedical Engineering Conference (CIBEC). Presented at the 2018 9th Cairo International Biomedical Engineering Conference (CIBEC), pp. 90–93. https://doi.org/10.1109/CIBEC.2018.8641762

Jagadish Kumar, S., Maheswaran, U., Jaikishan, G., Divagar, B., 2021. Melanoma Classification using XGB Classifier and EfficientNet, in: 2021 International Conference on Intelligent Technologies (CONIT). Presented at the 2021 International Conference on Intelligent Technologies (CONIT), pp. 1–4. https://doi.org/10.1109/CONIT51480.2021.9498424

Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D., 2019. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 17, 195. https://doi.org/10.1186/s12916-019-1426-2

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D.D., Chen, M., 2014. Medical image classification with convolutional neural network, in: 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV). Presented at the 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), pp. 844–848. https://doi.org/10.1109/ICARCV.2014.7064414

Mazzanti, S., 2021. SHAP explained the way I wish someone explained it to me [WWW Document]. Medium. URL https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30 (accessed 12.13.22).

Reza, M.T., Ahmed, F., Sharar, S., Rasel, A.A., 2021. Interpretable Retinal Disease Classification from OCT Images Using Deep Neural Network and Explainable AI, in: 2021 International Conference on Electronics, Communications and Information Technology (ICECIT). Presented at the 2021 International Conference on Electronics, Communications, and Information Technology (ICECIT), pp. 1–4. https://doi.org/10.1109/ICECIT54077.2021.9641066

Shakil, M.H., Rabiul Alam, Md.G., 2022. Toxic Voice Classification Implementing CNN-LSTM & Employing Supervised Machine Learning Algorithms Through Explainable AI-SHAP, in: 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET). Presented at the 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), pp. 1–6. https://doi.org/10.1109/IICAIET55139.2022.9936775

Tjoa, E., Guan, C., 2021. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Trans. Neural Netw. Learn. Syst. 32, 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314