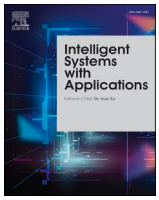
Intelligent Systems with Applications 20 (2023) 200275

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications



Skin cancer classification using explainable artificial intelligence on pre-extracted image features

Tarek Khater [a], Sam Ansari [a], Soliman Mahmoud [a], Abir Hussain [a,b,*], Hissam Tawfik [a]

[a] Electrical Engineering Department University of Sharjah Sharjah United Arab Emirates
[b] School of Computer Science and Mathematics Faculty of Engineering Liverpool John Moores University Liverpool L3 3AF UK

ARTICLE INFO

Keywords:
Artificial intelligence
Classification
Preprocessed images
Skin cancer
SHAP

ABSTRACT

Skin cancer is the most common type of cancer worldwide, affecting a large population recently. To date, various machine learning techniques exploiting skin images have been applied directly to skin cancer classification, showing promising results in improving diagnostic accuracy. This study aims to develop a machine learning-based model capable of accurately classifying skin cancer by utilizing extracted features from preprocessed images in the publicly available PH^2 dataset. Preprocessed features are known to provide more significant information than raw image data, as they capture specific characteristics of the images that are relevant to the classification task. The proposed model of this study can identify the most pertinent information in the images more accurately, thereby improving the performance and interpretability of the machine learning classification. Our simulation results illustrate that employing XG-boost yields an accuracy of 94% and an area under the curve value of 0.9947, further indicating that the proposed technique effectively distinguishes between non-melanoma and melanoma skin cancer. Explainable artificial intelligence provides some explanations by leveraging model-agnostic methods such as partial dependence plot, permutation importance, and SHAP. Moreover, the explainable artificial intelligence results show that asymmetry and pigment network features are the most important feature in the classification of skin cancer. These specific characteristics emerge as the most influential factors in distinguishing between different types of skin cancer.

1. Introduction

Skin cancer is a highly prevalent disease that affects a great number of people globally. It is distinguished by aberrant cell growth within the skin, which can result in tumor formation. The most frequent types of skin cancer are melanoma, basal cell carcinoma, and squamous cell carcinoma. These tumors can develop from several types of skin cells, including basal cells, squamous cells, and melanocytes (Gloster & Neal, 2006). The etiology of skin cancer is complicated, involving a combination of genetic, environmental, and lifestyle factors. Despite advances in cancer research and treatment, skin cancer remains a major public health concern, underscoring the importance of ongoing research and public awareness initiatives to prevent and manage this disease. Skin cancer can be caused by various factors, such as excessive exposure to ultraviolet (UV) radiation, genetics, and environmental factors.

According to the American Cancer Society, skin cancer accounts for approximately one-third of all diagnosed cancer cases in the United States. Skin cancer comes in two primary varieties: non-melanoma, which is more prevalent, and the uncommon melanoma variety. Because it spreads more quickly, melanoma is more hazardous and can be fatal if caught in its later stages (Mukherjee et al., 2019). With 1.5 million new cases in 2020, skin cancers are the most prevalent type of cancer diagnosed globally. A projected 325,000 new melanoma cases were diagnosed in 2020, and 57,000 people globally passed away from the condition (International agency for research on cacner 2022). In the United States in 2022, there were 97,920 new cases of melanoma in situ of the skin (Siegel et al., 2022). According to the American Cancer Society (Gomaa et al., 2022), three million skin cancer cases could be avoided each year if people were more aware of the risk factors associated with sun exposure and prevention.

Between first January and 31th December 2019, the UAE National Cancer Registry (UAE-NCR) received reports of 4633 newly diagnosed cancer cases, including both malignant and benign cancers (Al-Shamsi, 2022). Early detection of cancer is crucial for managing and curing diseases at its earliest stages, this applies to skin cancer, which can be identified, treated, and cured similarly to other illnesses. Skin

* Corresponding author.
E-mail address: abir.hussain@sharjah.ac.ae (A. Hussain).

https://doi.org/10.1016/j.iswa.2023.200275
Received 29 May 2023; Received in revised form 21 August 2023; Accepted 29 August 2023
Available online 3 September 2023

irregularity is referred to as a lesion. Cancerous, allergic, and other types of skin lesions are possible. The healthiest of these are skin blemishes that can cause cancer. These malignant lesions might be fatal in some cases. Among malignant lesions, melanoma is thought to have the highest fatality rate at 8%. The frequency of melanoma cases is rising daily (Jiang et al., 2020).

A non-invasive (Pham et al., 2019) diagnostic technique called dermoscopy enables practitioners to analyze the morphological structure of pigmented skin lesions. By analyzing the images produced by the dermoscopy instrument, melanoma is diagnosed. Dermatologists typically use the asymmetrical shape, border, color, and diameter (ABCD) criterion to diagnose melanoma through these photos. Based on the expertise and judgment of the relevant doctors, ABCD is a highly subjective evaluation (Pham et al., 2019). The ABCD rule-based and computer-assisted approaches can enhance melanoma diagnosis. Typically, ABCD systems have independent components for picture segmentation, feature extraction, and classification, as shown in Fig. 1. The classification accuracy of skin lesions is critical in the diagnosis and treatment of skin cancer. Traditional methods of skin cancer classification rely on human visual inspection, which can be prone to subjective interpretation and variability.

In recent years, machine learning (ML) has emerged as a promising approach to developing automated and objective skin cancer classification models. The abundance of skin lesion images, as well as the availability of advanced image processing techniques, has led to an increase in the application of ML in skin cancer classification. By training ML algorithms on large datasets of skin lesion images or even the feature extracted from the images, it is possible to develop models that can accurately classify skin cancer lesions into benign or malignant categories. With respect to image classification, the convolution neural network is the preferred algorithm to train ML models. CNNs are comprised of various types of layers including convolutional, pooling, and fully connected layers. Other algorithms can be utilized in the case of numerical or categorical data. These algorithms include decision trees, XG-boost, random forrest, and support vector machine (SVM).

A decision tree (Mahesh, 2020) is a graph that displays possibilities and their outcomes as a tree. The edges of the graph indicate the conditions or rules for making decisions, whereas the nodes in the graph represent an event or a choice. There are nodes and branches in every tree. Each node represents a set of characteristics that needs to be categorized, and each branch indicates a possible value for the node. Ensembles are methods that combine many ML algorithms to create a more powerful model or algorithm. For instance, the extreme gradient boosting (XG-boost) algorithm (Torlay et al., 2017) has many advantages such as dealing with missing values, and data scale requirements, suggesting a gradient boosting technique variation that is computationally effective and providing satisfactory results in ML performance.

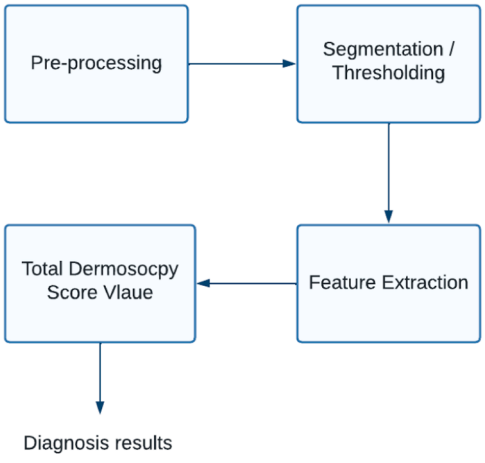


Fig. 1. Automated dermoscopy analysis for the detection of melanoma.

Random forest (Müller & Guido, 2016) is just a group of decision trees, each of which differs somewhat from the others. The theory behind random forests is that while each tree may make somewhat accurate predictions, it might overfit some portions of the data. By averaging the outcomes of several trees, we can lessen the degree of overfitting if they are all successful and overfit in various ways.

Different from other approaches that utilize images to classify skin cancer, our proposed approach deploys various ML algorithms to classify directly skin cancer as melanoma or non-melanoma based on the extracted features from the preprocessed images.

The remainder of this paper is organized as follows. Section 2 discusses the explainable artificial intelligence concept. Section 3 represents related works about skin cancer classification using different ML algorithms. Section 5 represents the case scenario and dataset. Accordingly, Section 6 discusses the results and discussion. Finally, significant conclusions are reported in Section 7.

2. Explainable artificial intelligence

Explainable Artificial intelligence (XAI) is (Gianfagna & Di Cecco, 2021) a collection of techniques and methods that can be used to explain the outcomes of the development of ML models in a way that is understandable to humans. There are two terms that should be mentioned which are interpretability and explainability. Interpretability (Thampi & Interpretable, 2022) comes down to comprehending cause and effect in an artificial intelligence (AI) system. It refers to how accurately we can predict what a model will predict given an input, how the model arrived at the forecast, how the prediction changes as the input or algorithmic parameters change, and lastly how accurately we can detect when the model has made a mistake.

Explainability (Thampi & Interpretable, 2022), meanwhile, goes beyond interpretability by assisting us in understanding in a way that is understandable to humans how and why a model arrived at a forecast. With the goal of reaching a far larger audience, it describes the internal workings of the system in straightforward words. In addition to using interpretability as a foundation, explainability also refers to other disciplines and topics including human-computer interaction (HCI), law, and ethics. The question is why XAI is needed and why it is so important. Relying solely on a single metric, such as classification accuracy, may not adequately capture and address the complexities of the situation at hand. Simply obtaining the value of accuracy without understanding the underlying factors that contribute to it renders the information incomplete and potentially useless. The three primary uses for ML models that frequently involve prediction and necessitate explainability are model debugging, model validation, and knowledge discovery.

XAI has two main approaches or methods which are the intrinsic approach which means that the internal parameters of the model are used to generate interpretations and the model agnostic approach which means that used when the model is a black box and we can't know the internal parameters. There are various types of explanations such as intrinsic which is known as post hoc, model-specific which is referred to as model-agnostic, and global or local explanations. In this paper, the type of ML model used is the model-agnostic model.

Model-agnostic methods have emerged as powerful techniques for generating explanations in machine learning, avoiding the need to rely on the internal workings of models that are often characterized as "opaque". By decoupling the explanation generation process from model-specific details, model-agnostic methods offer a versatile approach for producing interpretable explanations in a wide range of ML applications. These methods include the permutation importance method, partial dependence plots (PDPs), and Shapley additive explanations. Permutation importance allows the detection of the most significant features. It is based on shuffling the values of a feature and repeating the prediction again while monitoring the error. If the error

gets worsens, this means that this feature is important and highly impacts the prediction. Consequently, the shuffling process leads to a deterioration in predictions as the significance of specific features increases.

PDP provides details on how these features are impacting the predictions. It is a plot that shows the functional relationship between one input or more and the output target. From the PDP we can see how the change in the prediction can be affected by the most important features. Shapley additive operations depend on Shapley values that provide explanations on specific instances as well as global explanations. So, it changes the direction of explanations from global to local explanations. We can determine which feature is more important to a given prediction using Shapley values. When we require an answer for a particular prediction and are less concerned with knowing the model's "typical" behavior, SHAP can be helpful. XAI has numerous applications in healthcare. For instance, XAI can be utilized in medical image analysis and clinical decision support by building explainable ML models to help in the early diagnosis stage. By examining the contribution of biomarkers or clinical characteristics to a particular disease outcome, SHAP has the potential to be applied in the field of healthcare.

3. Related works

Ozkan & Koklu (2017) pre-classified the skin lesions into normal, abnormal, and melanoma. They designed a machine-learning model to support the decision of the doctors. Skin lesions based on dermoscopic pictures from PH^2 datasets (Mendon¸ca et al., 2015) are the study's main focus. Four different ML techniques, artificial neural network (ANN), SVM, K-nearest neighbor (KNN), and decision tree, are used to achieve this goal. For ANN, SVM, KNN, and decision tree, the simulation results showed an accuracy of 92.50%, 89.50%, 82.00%, and 90.00%, respectively. Alkarakatly et al. (2020) designed a 5 layers convolution neural network (CNN) classifier of skin lesions which is melanoma or nevus based on the PH^2 dataset. The performance of the model was evaluated by classification accuracy, sensitivity, specificity, and the area under the curve (AUC). On the test set, it scored 100% AUC, 94% sensitivity, 97% specificity, and around 95% accuracy. In another study (Mukherjee et al., 2019), Soumen Mukherjee et al used DERMOFIT and MEDNODE, representing malignant lesion image datasets, employed independently and jointly to assess the effectiveness of their proposed CNN presented as CNN malignant lesion detection (CMLD). When these datasets were used separately, the accuracy was 90.58% for DERMOFIT and 90.14% for MEDNODE.

When they were combined, it achieved 83.07% accuracy.

Shahsavari et al. (2022) designed a new computer-aided method called the En-semble of Deep (SLDED) model in order to detect skin lesions. By using the ISIC archive database, which had 4668 skin lesion images for lesion localization, they used a modified faster Regions with CNN networks (R-CNN) with deep learning model (VGGNet) feature extractor and achieved a mean average precision (mAP) of 0.96. They assess the experimental classification outcomes on 934 and 200 images using test data from ISIC (Gutman et al., 2016)and PH^2 (Mendon¸ca et al., 2015). For ISIC and PH^2 test data, they achieved an average accuracy of 97.1% and 96%, precision of 87.1% and 90.2%, AUC of 98.6% and 98.1%, and recall of 86.7% and 85.4%, respectively.

Jiang et al. (2020) proposed an end-to-end framework called Channel & Spatial Attention Residual Module (CSARM-CNN) model, which can segment skin lesions effectively and automatically. By using spatial pyramid pooling, multiscale input images were obtained. In order to sum the model's overall loss, a weighted cross-entropy loss function was applied to each side of the output layer. The authors conducted their evaluations using the two publicly available standard datasets ISIC 2017 and PH^2, and their findings were competitive in terms of accuracy and specificity, with 94.96% and 95.23% accuracy, and 99.03% and 99.45% specificity, respectively.

Kumar & Vatsa (2022) reviewed and analyzed two deep neural-based classification algorithms including a convolution neural network and recurrent neural network as well as a decision tree-based algorithm (XG-Boost) on the ISIC dataset. The authors attempted to determine which one has the best categorization performance metric. Loss, precision, accuracy, recall, ROC, and F1 score are used to benchmark how well algorithms work. They indicated that the VGG16 architecture performed the best for CNN, with an accuracy of 89.6%. The RNN's bidirectional architecture is also superior to the other RNN architectures (accuracy: 95.96%). The XG-Boost approach has a 97.22% accuracy rate.

Hosny et al. (2018) proposed an automated method for classifying skin lesions. This approach makes use of deep transfer learning, in which the final layer of AlexNet is replaced with a softmax to categorize three different lesions. The PH^2 dataset is used to train and evaluate the suggested model. The performance of the suggested technique is assessed using quantitative measures of accuracy, sensitivity, specificity, and precision, with achieving values of 98.61%, 98.33%, 98.93%, and 97.73%, respectively. In another study, Iftiaz A. Alfi (Alfi et al., 2022) presented an interpretable approach for the ensemble stacking of ML models and deep learning for the non-invasive diagnosis of melanoma skin cancer.

Logistic regression, random forest, SVM, XG-boost, and KNN are trained using manually extracted features. Transfer learning was carried out using pretrained deep learning models (MobileNet, ResNet50, Xception, DenseNet121, and ResNet50V2) using ImageNet data. Ensembled ML models with deep learning architectures are performed and evaluatedThey determined the most accurate model for categorizing skin lesions by calculating accuracy, Cohen's kappa, F1-score, ROC curves, and confusion matrix.

The Review of previous research reveals that most of the studies use skin images directly for the application of ML classification rather than features extracted from the preprocessed images. Hence, The aim of this work is to focus on classifying different skin cancer categories based on preprocessed images. This could allow potentially more interpretable ML models to be used to investigate the effect of the extracted characteristics on the categorization or prediction of melanoma or non-melanoma skin cancer.

4. Explainable machine learning related works

Singh et al. (2020) used Kernel Shapley Additive explanations (SHAP) and GradCAM to compare 30 CNN models. It was demonstrated that even very accurate models occasionally concentrated on features that weren't crucial for the diagnosis. The attribution maps of both methods demonstrated that there were variations in the models' explanations for similar accuracy. This demonstrated that various neural network topologies have the propensity to learn various features. Van Molle et al. (2018) exhibited the features of a convolutional neural network (CNN) for the classification of skin lesions. The layers were seen to be examining risk factors including lighter skin tone or a pinkish texture as well as markers like lesion boundaries and color irregularity. However, insignificant traits like hair and artifacts were also learned, indicating some degree of overfitting. By using XAI techniques, Dindorf et al. (2020) looked at how different input representations affected a trained model's accuracy, interpretability, and clinical relevance. Using an inertial measuring unit (IMU)-based device, the gait of 27 healthy patients and 20 subjects who had had total hip arthroplasty (THA) was captured. For categorization, three distinct input representations were used. The model interpretation was carried out using Local Interpretable Model-Agnostic Explanations (LIME). The features that were automatically extracted provided the greatest accuracy.

In another study, Binder et al. (2021) provide an understandable machine-learning approach for the combined profiling of morphological, molecular, and clinical data from breast cancer histology. First, their method enables accurate heatmap representations of the classifier decisions and the robust detection of cancer cells and tumor-infiltrating

lymphocytes in histology images. Second, histology can be used to predict molecular characteristics such as DNA methylation, gene expression, copy number variations, somatic mutations, and proteins. Balanced accuracy of molecular predictions is up to 78%, whereas accuracy for patient subgroups can reach over 95%. Last but not least, their defensible AI strategy enables the evaluation of the relationship between morphological and molecular cancer features.

During the Covid-19 pandemic, Magunia et al. (Suri et al., 2020) developed an ML-based model to classify patient risk and predict ICU outcomes, based on retrospective and prospective clinical data. Prediction accuracy and readability were assessed for ML methods. The Explainable Boosting Machine strategy was determined to be the best course of action. As a result, it was determined that the model for predicting the general outcome of the ICU was more accurate in predicting "survival". Age, thrombotic and inflammatory activity, and the degree of ARDS (Acute respiratory distress syndrome) at ICU admission were found to be indicators of ICU survival. Qu et al. (2022) tried to predict the occurrence of congenital heart diseases using innovative machine-learning techniques. cardiac hospitals.

ROC curves and the explainable boosting machine (EBM) for AUC prediction were used to evaluate the model's performance. The most effective predictors were chosen based on their contributions and predicting abilities. The most significant predictors have thresholds determined for them. The model achieved an AUC of 76% (69-83%), and total accuracy, sensitivity, and specificity were 0.65, 0.74, and 0.65, respectively. The Total accuracy, specificity, and sensitivity were 0.65, 0.65, and 0.74, respectively.

Pavan et al. (Magesh et al., 2020) proposed an ML model that accurately categorizes each given DaTSCAN as having Parkinson's disease or not, in addition to offering a logical explanation for the prediction. Visual cues produced utilizing Local Interpretable Model-Agnostic Explainer (LIME) techniques are used in this type of reasoning. Transfer learning was used to train DaTSCANs on a CNN (VGG16) from the Parkinson's Progression Markers Initiative database, and the resulting models had 95.2% accuracy, 97.5% sensitivity, and 90.9% specificity. This study uses visual supe rpixels on the DaTSCANs to identify PD from non- PD using LIME explanations since model interpretability is crucial, especially in the healthcare industry.

Yoo et al. (2020) developed an interpretable multiclass ML model that selects the laser surgery option on the expert level. The Shapley Additive ex-Planation technique was adopted to explain the output of the XG-Boost model. When tested on the internal and external validation datasets, the multiclass XGBoost model showed an accuracy of 81.0% and 78.9%, respectively. The results of the Shapley Additive ex-Planations explanations were in line with what ophthalmologists already knew. The one-versus-one and one-versus-rest XGBoost classifiers' explanations were successful in making users of the multi-categorical classification problem understandable.

5. Materials and methods

In this section, the PH^2 dataset, the definitions of the features, and the methodology of the skin cancer classification are described and presented.

5.1. $PH2^2$ dataset

Researchers from the Technical Universities of Porto and Lisbon created this data collection in the dermatology department of Pedro Hispano Hospital (Mendon¸ca et al., 2015). The 200 dermoscopy images in the PH^2 dataset have a resolution of 768×560. The dataset contains seven input features and one output feature. Firstly, the asymmetrical feature knowing that asymmetry in skin lesions is a reliable sign of malignant melanoma. This means that the shape of one half does not resemble the other half (Ali et al., 2020). Secondly, the pigment network has a linear shape and looks like hair artifacts (Alfed et al., 2015) and it has brown lines. Thirdly, the Dots/Globules feature is used to describe black, brown, round to oval, variously sized objects that were dispersed either regularly or erratically within a melanocytic lesion (Xu et al., 2009). Fourthly, the streaks which are linear extensions of pigment at the edge of a lesion as radially structured linear structures in the direction of growth (Sadeghi et al., 2013). Then, the regression areas feature which is identified in the dermoscopic lesion image by the presence of white and grey-blue patches (Bassoli et al., 2011). The structureless zones of confluent blue pigment with a ground-glass blur are known as the "blue-white veil regions" (Madooei et al., 2013). Finally, the number of colors in skin cancer. Some cancers have multiple colors such as white, black, red, Dark-Brown, Light- Brown, and Blue Gray.

5.2. Methods

The typical direction or methodology to classify skin cancer is to use the skin images as input to the ML model. The images are passed to a CNN which identifies the objects in the images, hence the model can learn to discriminate between melanoma and non-melanoma skin cancer. In this study, a preprocessing step was performed on the images followed by the extraction of features from these preprocessed images. By utilizing these features, the ML model can precisely recognize the crucial information present in the images, leading to an enhancement in the accuracy and comprehensibility of the model. We used classical ML algorithms including KNN, XG-boost, decision tree, and random forest. Fig. 2 depicts the process of building the ML model on pre-extracted image features.

Finally, in order to provide insight into the potential diagnosis of breast cancer, the model's output and predictions were examined and interpreted using XAI.

6. Results and discussion

In this section, some descriptive analysis is described in addition to the ML model results.

6.1. Descriptive analysis

In terms of data description, Fig. 3a shows the histogram for the clinical diagnosis which is the output target while Figs. 3b through 3h represent the histogram of each input feature. There are three classes: typical nevus, atypical nevus, and melanoma. Further descriptive statistical analysis had been performed on the dataset. For instance, the correlation matrix is used to find the relation between the inputs themselves and the inputs with the outputs. It is clear from the correlation matrix that there are no two features that are highly correlated. The correlation matrix is shown in Fig. 4.

6.2. Feature importance

The input feature importance was estimated using chi-square. Fig. 5 shows that the asymmetry and pigment network features have the highest chi-square score which means that these features have the strongest relationship with the target variable among all the others features being considered. This means that it is likely to have the greatest impact on the accuracy of the model and should be given priority in the feature selection process.

6.3. ML model performance

Multiple ML algorithms were utilized to train the model, including XGBoost, decision tree, random forest, and KNN. The performance of the model was evaluated based on precision, recall, and f-score. The classification simulation results for each algorithm are presented in Fig. 6. The XGBoost and decision tree algorithms yielded the highest accuracy,

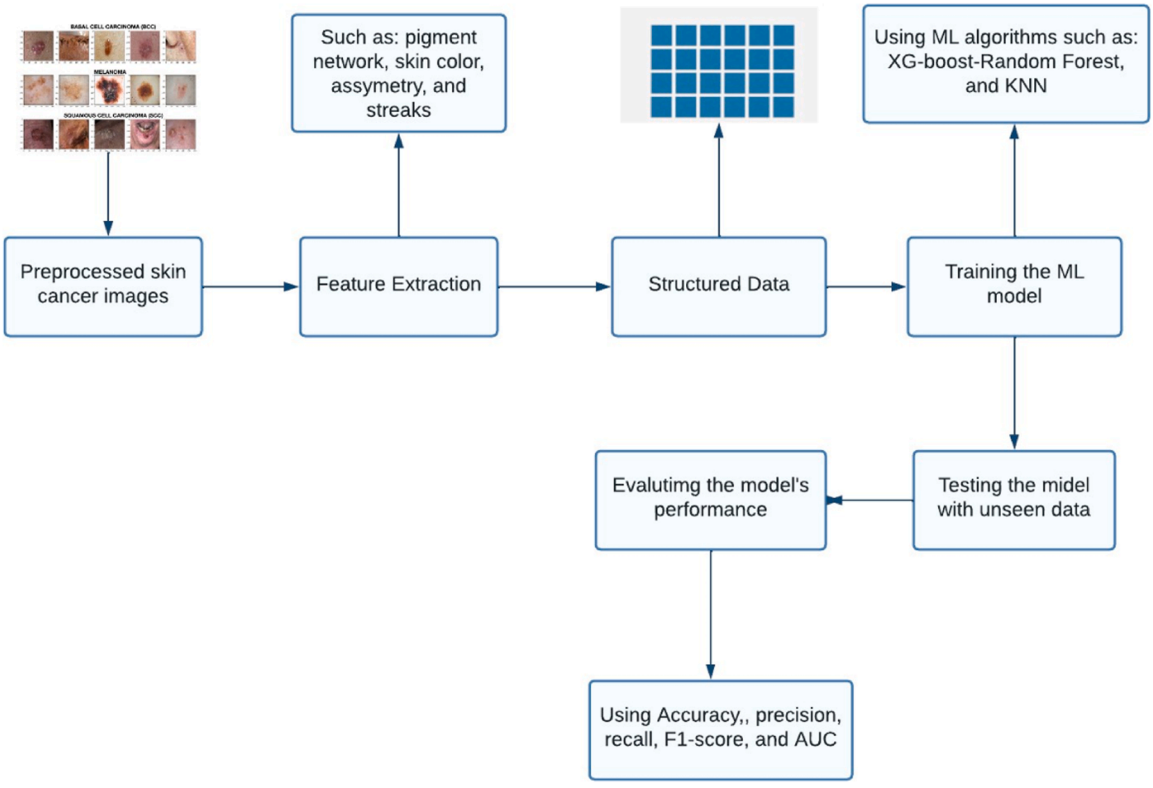


Fig. 2. Complete process of developing ML model.

achieving 94%. Prior research in this field has been limited, with only one paper (Ozkan & Koklu, 2017) utilizing preprocessed image features to achieve 92% and 90% accuracy using KNN and decision tree algorithms, respectively. In this study, we attained higher accuracy when using a decision tree, and equivalent accuracy using a KNN, as compared to (Ozkan & Koklu, 2017).

Moreover, we trained the model using XGBoost, which resulted in accuracy surpassing that achieved using the KNN in (Ozkan & Koklu, 2017). To assure that our model gives better performance, AUC for ROC had been estimated. AUC is the ability of the model to differentiate between the positive classes and the negative ones. Due to it being a multiclassification problem, the One vs All technique is used. Fig. 7 shows the values of the AUC for each algorithm. Fig. 8 shows the ROC curve which is the relation between the true positive rate and the false positive rate. It is clear from the curve that the highest AUC value is for the typical nevus class which indicates that the model has a high degree of discriminatory power in classifying this level compared to the other skin cancer categories.

6.4. XAI results

In order to understand why the ML model is making the predictions that it is, it is essential to explain and analyze the results once the model's performance has been evaluated. This involves understanding the relationships between the features and the goal variable, identifying any pertinent patterns or trends in the data, and identifying the features that are essential for the model's predictions.

Our best-performing is achieved using the XG-boost algorithm.

6.4.1. Permutation importance results

The feature's importance is determined by evaluating the change in the model prediction error after the permutation. When a feature is used by the model to make predictions in this particular situation, it is said to be "essential" if changing its values results in an increase in model error. If changing a feature's values results in the same model error as when leaving it alone for the forecast, the feature is said to be "unimportant." Fig. 9 shows the ranking of the features which reveals that the pigment network feature is the most important feature for the ML model to classify breast cancer.

6.4.2. Partial dependence plot results

The PDP displays a predictor variable's marginal effect, which is its average influence throughout the entire dataset, on the target variable. This graph can be used to spot interactions between predictors as well as non-linear correlations between the predictor and the target variables. The PDP illustrates how, despite keeping all other inputs constant, the anticipated outcome varies as the binary input shifts from one level to the next. Fig. 10 depicts that when the asymmetry feature changes from 0 to 1 and from 1 to 2, the prediction of the melanoma class increases. On the other hand, when the asymmetry feature changes from 0 to 1 and from 1 to 2, the prediction of the nevus whether it is a common or atypical class decreases.

6.4.3. Shapley results

SHAP can produce local explanations by localizing the model using a smaller model and perturbing the input data to observe how the output changes. One instance is chosen from the dataset to see the impact of the features on the corresponding outcome. Fig. 11 shows the effect of the input features on the common nevus class. It is clear that the pigment network when equalling 0 pushes the model towards the left side to decrease the prediction value. Fig. 12 represents the local SHAP plot for the atypical nevus class. It is shown that when the pigment network feature equals 1 which has the highest length, the prediction value for the atypical nevus class is increased. Fig. 13 shows that when the asymmetry feature is 2 and the number of colors is 5, the prediction of melanoma increases.

Shapley values that offer explanations for specific occurrences rather than just general ones are necessary for Shapley additive operations

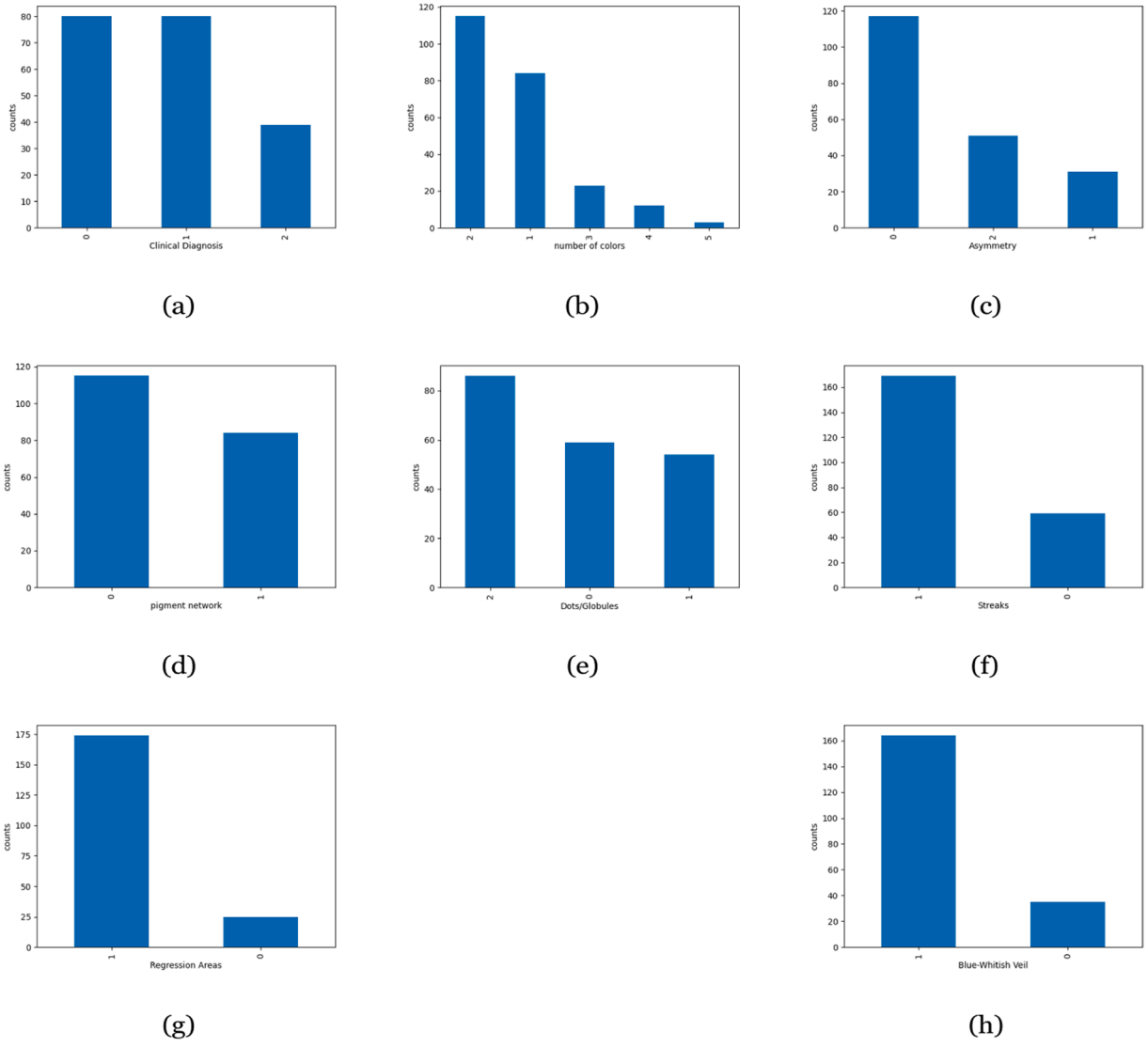


Fig. 3. Histogram plots of (a) Clinical diagnosis histogram in which 0 represents common Nevus, 1 represents Atypical Nevus, 2 represents melanoma, (b) Number of colors histogram, (c) Asymmetry histogram in which 0 is fully symmetric, 1 is symmetric in 1 axis, and 2 is fully asymmetric, (d) Pigment network histogram in which 0 represents atypical and 1 represents typical, (e) Dots/Globules histogram which 0 is atypical, 1 is typical, and 2 is absent, (f) Streaks histogram in which 1 is absent and 0 is present, (g) Regression areas which 1 is absent and 0 is present, (h) Blue-whitish veil histogram in which 1 is absent and 0 is present.

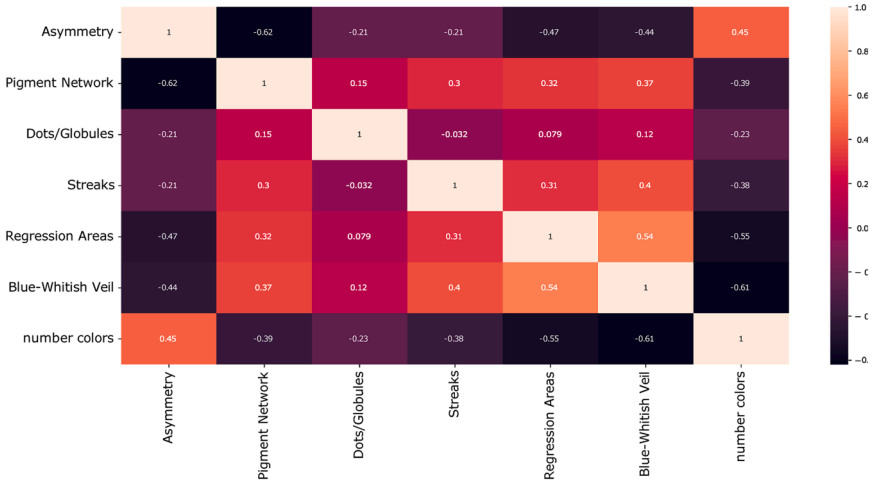


Fig. 4. The correlation matrix.

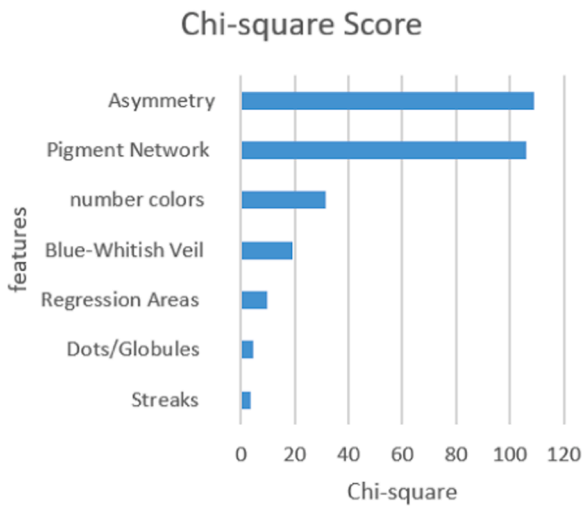


Fig. 5. Feature importance using chi-square scores.

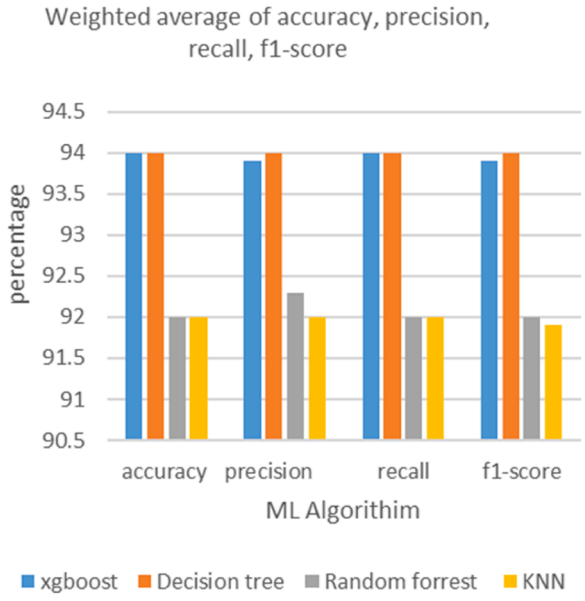


Fig. 6. The weighted average of accuracy, precision, recall, and f1-score.

(Gianfagna & Di Cecco, 2021). Using Shapley values, we may determine which feature is more crucial for a specific prediction. SHAP can be useful when we need an answer for a specific forecast and are less interested with understanding the model's "typical" behavior. By calculating the contribution of each characteristic to the prediction, SHAP (Molnar, 2023) seeks to explain the prediction of an instance x.

To provide a global explanation for the model, a summary SHAP plot is shown in Figs. 14 and 15 for common nevus and melanoma classes, respectively. Fig. 14 reveals that the pigment network is the most contributing feature in the prediction of the common nevus class. Notably, when the pigment network has a high value which is basically 2, it affects the ML prediction of the common nevus positively and vice versa. The SHAP plot in Fig. 15 depicts that the asymmetry feature has the most contribution to the prediction of the melanoma class and also has a positive impact on the prediction.

To summarize the SHAP results, the asymmetry feature and the pigment network play an important role in the prediction of melanoma and nevus of skin, respectively.

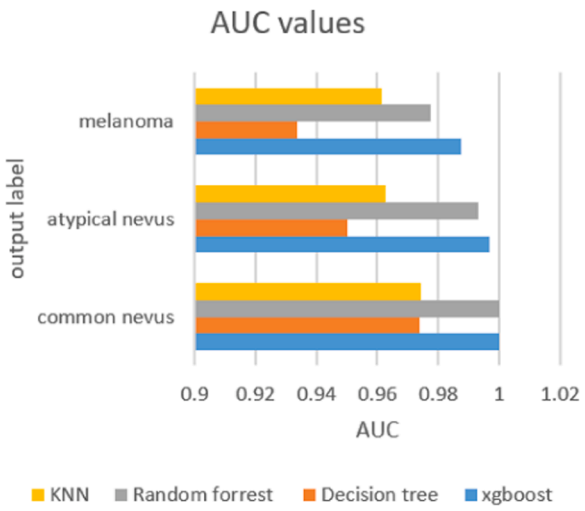


Fig. 7. The area under the curve of each algorithm.

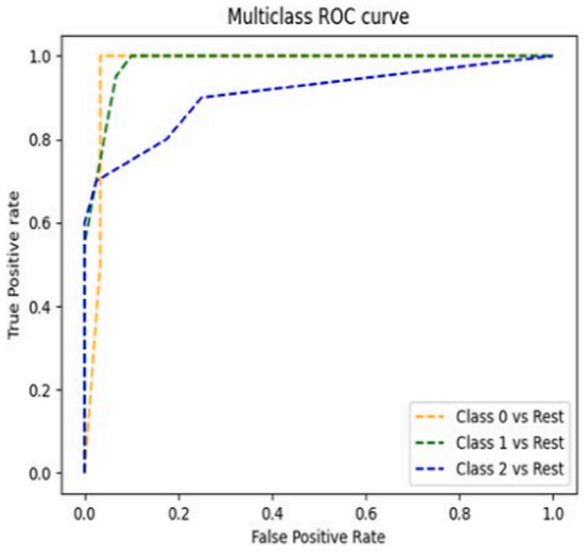


Fig. 8. ROC curve.

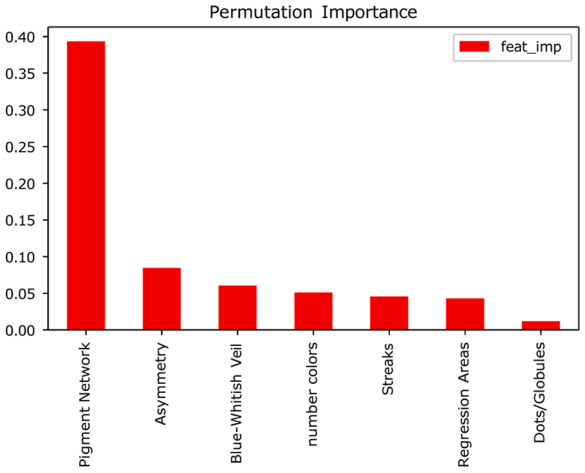


Fig. 9. Permutation importance plot.

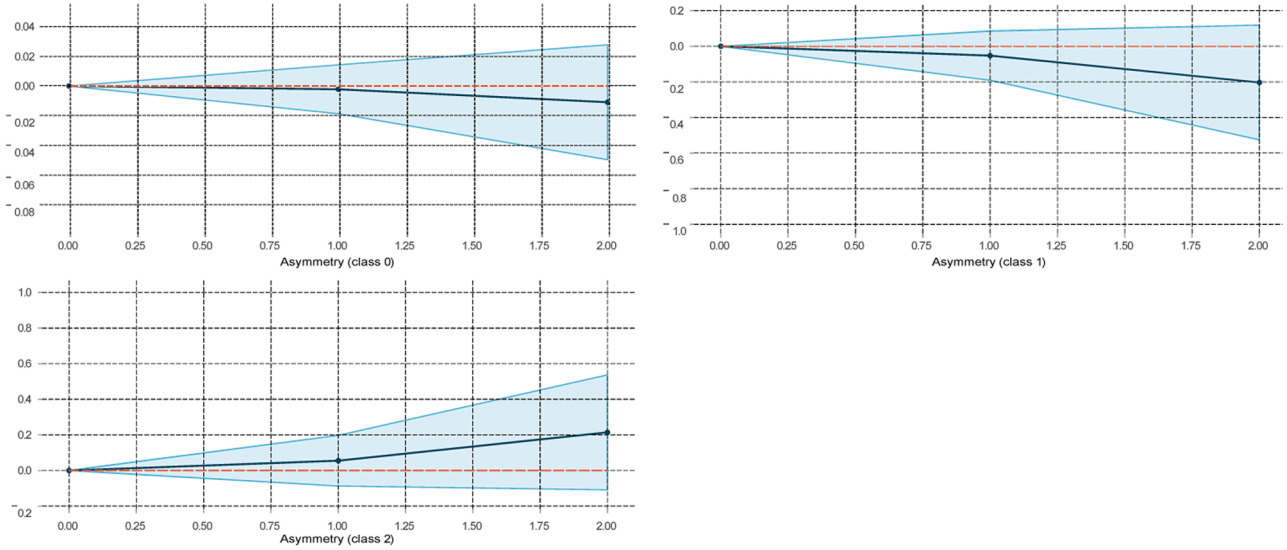


Fig. 10. PDP plot for the asymmetry.

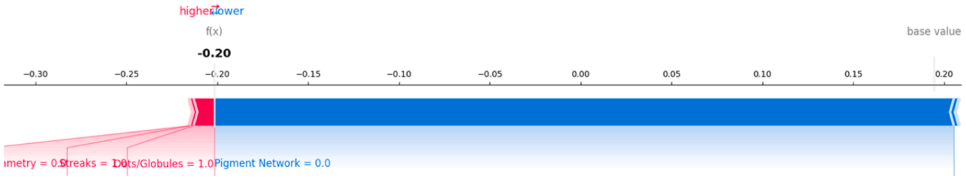


Fig. 11. SHAP plot for the common nevus class.

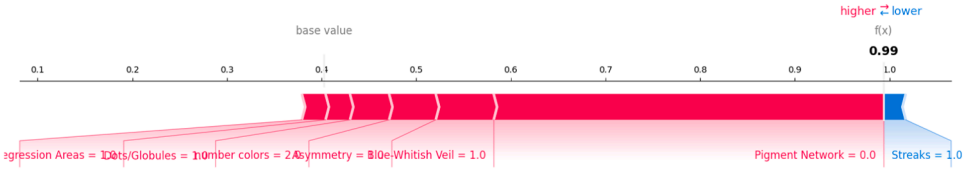


Fig. 12. SHAP plot for the Atypical nevus class.

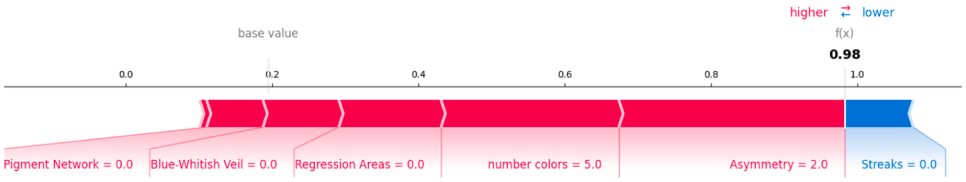


Fig. 13. SHAP plot for the melanoma class.

7. Conclusion & future work

The utilization of preprocessed features has been shown to enhance the effectiveness, precision, and comprehensibility of ML models in image classification tasks, including the classification of skin cancer. By emphasizing the most significant information embedded in the images and reducing the dimensionality of the input data, ML models can be trained more effectively, offering an additional layer of precision and objectivity in the diagnosis of skin cancer. Feature importance is performed using the chi-square method showing that the asymmetry and pigment network are the most important features. This research employs multiple ML algorithms, such as XG-boost, decision trees, random forest, and KNN, to train our model. The simulation results indicate an accuracy of 94% for both XG-boost and decision tree, further designating the significant superiority of the proposed framework. XAI is utilized to provide explanations, local or global, of the model results, facilitating a better understanding of the model behavior. According to the XAI analysis of skin cancer classification, asymmetry and pigment network traits are among the most crucial characteristics in evaluating whether a skin lesion is malignant or benign. These characteristics are important markers that dermatologists and AI models may leverage to assess skin lesions, even if they are not directly related. The results of the XAI study support pre-existing medical understanding of the visual features of malignant skin lesions and demonstrate the potential for AI technology to assist in the detection and management of skin cancer. Other model-agnostic methods, such as LIME, can be employed in the future in order

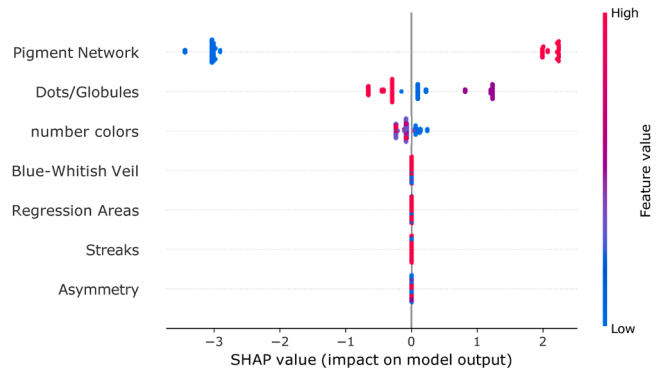


Fig. 14. SHAP summary plot for the common nevus class.

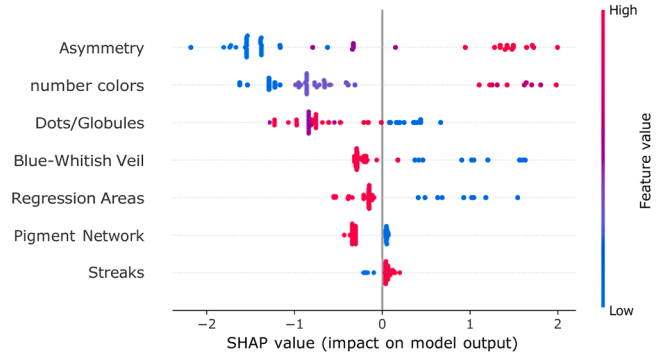


Fig. 15. SHAP summary plot for the melanoma class.

to produce more local explanations for each patient individually.

Credit Author Statement

Tarek Khater: performed conceptualized, methodology, simulation, visualization, data analysis and writing. **Sam Ansari:** performed data analysis and proof reading. **Soliman Mahmoud:** performed writing-review and editing. **Abir Hussain:** performed writing- review and editing. **Hissam Tawfik:** performed, supervision, and writing- review and editing

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Al-Shamsi, H. O. (2022). The state of cancer care in the united arab emirates in 2022. *Clinics and Practice, 12*(6), 955–985.

Alfed, N., Khelifi, F., Bouridane, A., & Seker, H. (2015). Pigment network-based skin cancer detection. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 7214–7217). IEEE.

Alfi, I. A., Rahman, M. M., Shorfuzzaman, M., & Nazir, A. (2022). A non-invasive interpretable diagnosis of melanoma skin cancer using deep learning and ensemble stacking of machine learning models. *Diagnostics, 12*(3), 726.

Ali, A.-R., Li, J., & O'Shea, S. J. (2020). Towards the automatic detection of skin lesion shape asymmetry, color variegation and diameter in dermoscopic images. *PLoS One, 15*(6), Article e0234352.

Alkarakatly, T., Eidhah, S., Al-Sarawani, M., Al-Sobhi, A., & Bilal, M. (2020). In *Skin lesions identification using deep convolutional neural network, in: 2019 International Conference on Advances in the Emerging Computing Technologies (AECT)* (pp. 1–5). IEEE.

Bassoli, S., Borsari, S., Ferrari, C., Giusti, F., Pellacani, G., Ponti, G., & Seidenari, S. (2011). Grey-blue regression in melanoma in situ—evaluation on 111 cases. *Journal of Skin Cancer*, 2011.

Binder, A., Bockmayr, M., Hägele, M., Wienert, S., Heim, D., Hellweg, K., Ishii, M., Stenzinger, A., Hocke, A., Denkert, C., et al. (2021). Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence, 3*(4), 355–366.

Dindorf, C., Teufl, W., Taetz, B., Bleser, G., & Fröhlich, M. (2020). Interpretability of input representations for gait classification in patients after total hip arthroplasty. *Sensors, 20*(16), 4385.

Gianfagna, L., & Di Cecco, A. (2021). *Explainable AI with python*. Springer.

Gloster, H. M., Jr, & Neal, K. (2006). Skin cancer in skin of color. *Journal of the American Academy of Dermatology, 55*(5), 741–760.

Gomaa, B., Houghton, R. F., Crocker, N., & Walsh-Buhi, E. R. (2022). Skin cancer narratives on Instagram: content analysis. *JMIR Infodemiology, 2*(1), e34940.

Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., & Halpern, A. (2016). *arXiv preprint*.

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)* (pp. 90–93). IEEE.

International agency for research on cacner (2022). URL https://www.iarc.who.int/cancer-type/skin-cancer/.

Jiang, Y., Cao, S., Tao, S., & Zhang, H. (2020). Skin lesion segmentation based on multi-scale attention convolutional neural network. *IEEE Access, 8*, 122811–122825.

Kumar, A., & Vatsa, A. (2022). Untangling classification methods for melanoma skin cancer. *Frontiers in Big Data, 5*.

Madooei, A., Drew, M. S., Sadeghi, M., & Atkins, M. S. (2013). Automatic detection of blue-white veil by discrete colour matching in dermoscopy images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference* (pp. 453–460). Springer. September 22-26, 2013Proceedings, Part III 16.

Magesh, P. R., Myloth, R. D., & Tom, R. J. (2020). An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery. *Computers in Biology and Medicine, 126*, Article 104041.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research, 9*, 381–386 [Internet].

Mendonça, T., Celebi, M., Mendonca, T., & Marques, J. (2015). Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy Image Analysis*.

Molnar, C., Interpretable machine learning: A Guide for Making Black Box Models Explainable. Accessed: May. 2, 2023. [Online]. Available: https://christophm.github.io/interpretable-ml-book.

Muller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. *O'Reilly Media, Inc*.

Mukherjee, S., Adhikari, A., & Roy, M. (2019). Malignant melanoma classification using cross-platform dataset with deep learning cnn architecture. In *Recent Trends in Signal and Image Processing: Proceedings of ISSIP 2018* (pp. 31–41). Springer.

Ozkan, I. A., & Koklu, M. (2017). Skin lesion classification using machine learning algorithms. *International Journal of Intelligent Systems and Applications in Engineering, 5*(4), 285–289.

Pham, H. N., Koay, C. Y., Chakraborty, T., Gupta, S., Tan, B. L., Wu, H., Vardhan, A., Nguyen, Q. H., Palaparthi, N. R., Nguyen, B. P., et al. (2019). Lesion segmentation and automated melanoma detection using deep convolutional neural networks and xgboost. In *2019 International Conference on System Science and Engineering (ICSSE)* (pp. 142–147). IEEE.

Qu, Y., Deng, X., Lin, S., Han, F., Chang, H. H., Ou, Y., Nie, Z., Mai, J., Wang, X., Gao, X., et al. (2022). Using innovative machine learning methods to screen and identify predictors of congenital heart diseases. *Frontiers in Cardiovascular Medicine, 8*, 2087.

Sadeghi, M., Lee, T. K., McLean, D., Lui, H., & Atkins, M. S. (2013). Detection and analysis of irregular streaks in dermoscopic images of skin lesions. *IEEE Transactions on Medical Imaging, 32*(5), 849–861.

Shahsavari, A., Khatibi, T., & Ranjbari, S. (2022). *Skin lesion detection using an ensemble of deep models: Slded* (pp. 1–20). Multimedia Tools and Applications.

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians, 72*(1), 7–33.

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging, 6*(6), 52.

Suri, J. S., Puvvula, A., Majhail, M., Biswas, M., Jamthikar, A. D., Saba, L., Faa, G., Singh, I. M., Oberleitner, R., Turk, M., et al. (2020). Integration of cardiovascular risk assessment with covid-19 using artificial intelligence. *Reviews in Cardiovascular Medicine, 21*(4), 541–560.

Thampi, A., & Interpretable, AI (2022). *Building explainable machine learning systems*. Simon and Schuster.

Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017). Machine learning–xgboost analysis of language networks to classify patients with epilepsy. *Brain Informatics, 4*(3), 159–169.

Van Molle, P., De Strooper, M., Verbelen, T., Vankeirsbilck, B., Simoens, P., & Dhoedt, B. (2018). Visualizing convolutional neural networks to improve decision support for skin lesion classification. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF*

2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1 (pp. 115–123). Springer.

Xu, J., Gupta, K., Stoecker, W. V., Krishnamurthy, Y., Rabinovitz, H. S., Bangert, A., Calcara, D., Oliviero, M., Malters, J. M., Drugge, R., et al. (2009). Analysis of globule types in malignant melanoma. *Archives of Dermatology, 145*(11), 1245–1251.

Yoo, T. K., Ryu, I. H., Choi, H., Kim, J. K., Lee, I. S., Kim, J. S., Lee, G., & Rim, T. H. (2020). Explainable machine learning approach as a tool to understand factors used to select the refractive surgery technique on the expert level. *Translational Vision Science & Technology, 9*(2), 8–8.