RESEARCH
www.commprac.com
ISSN 1462 2815

EXPLAINABLE AI-DRIVEN MEDICAL DIAGNOSIS: A NOVEL METHOD FOR CATEGORIZING SKIN LESIONS WHILE TAKING INTO ACCOUNT SARS-CoV-2

Krishna Mohan Pandey [1], and Dev Baloni [2]

[1,2] Department of Computer Science and Engineering,
Quantum University, Roorkee, Uttarakhand, India.
Email: [1]krish.mp81@gmail.com, [2]devbaloni1982@gmail.com

DOI: 10.5281/zenodo.10976681

Abstract

Early-stage skin lesions, especially those linked to different types of skin cancer, present challenges in precise classification because of morphological resemblances. Traditional deep learning techniques prove to be effective; however, their lack of transparency and interpretability hinders acceptance among medical professionals. This research introduces a new method that combines Ensemble Max Voting with Explainable Artificial Intelligence (XAI) strategies to develop a clear and efficient system for classifying skin lesions, taking into account implications related to SARS-CoV-2. By utilizing Max Voting to amalgamate the collective intelligence of various algorithms, this model enhances classification accuracy and robustness across different types of lesions. The incorporation of XAI methods provides easy-to-understand, visually interpretable justifications for the model's decisions, bridging the gap between machine learning and human comprehension. Leveraging the Interna- tional Skin Imaging Collaboration (ISIC) dataset, our model accurately identifies eight categories of skin lesions, achieving impressive performance metrics: accuracy (94.47%), precision (93.57%), recall (94.01%), and F1 score (94.45%). The merger of Ensemble Max Voting with XAI presents a promising reso- lution for dependable and transparent skin lesion classification, holding potential for practical clinical application and further exploration. Furthermore, predictions are scrutinized using the LIME and GradCAM framework, offering visual explanations that adhere to established standards and enhance usability in clinical environments, while also considering factors related to SARS-CoV-2.

Index Terms: Ensemble Deep learning, Explainable Artificial Intelligence (XAI), Computer-Aided Diagnosis, Transfer Learn- ing, Skin Lesion Classification.

I. INTRODUCTION

A Skin lesions are indeed areas of skin that appear different from the surrounding tissue or patch on the skin, which may be indicative of underlying health conditions, including various forms of skin cancer. These irregularities can manifest in diverse ways, such as moles, lumps, discolorations, or ulcers. These can manifest in various forms, such as: Benign Lesions: These include moles, warts, and seborrheic keratoses, which are generally harmless but may sometimes cause discomfort or aesthetic concerns. Malignant Lesions: These malignant formations, including melanoma, basal cell carcinoma, and squamous cell carcinoma, pose significant risks if not identified and addressed promptly. Inflammatory and Infectious Lesions: Conditions like psoriasis, eczema, and fungal infections can lead to lesions that indicate underlying health issues requiring medical attention [1].

The early detection of skin lesions, particularly malignant types, is vital as Early-stage skin cancer is typically more treat- able, with a wider range of therapeutic options available [2]. Timely intervention can prevent metastasis, where malignant cells from the primary cancerous site metastasize to disparate regions within the organism and greatly improve survival rates. Treating skin cancer at an advanced stage often requires more aggressive and invasive procedures. Early detection allows for simpler and less burdensome treatments, enhancing the quality of life for patients. Early

diagnosis and treatment generally reduce the overall cost of care, as late-stage interventions can be significantly more expensive and prolonged [3]. Prompt identification and accurate categorization of skin lesions are necessary for timely medical diagnosis, minimizing the risk of progression and improving patient outcomes. However, due to the wide range of appearances and subtle differences among skin lesion types, the task of accurate classification remains a complex challenge. In summary, skin lesions are complex and multifaceted phenomena, with implications that range from benign discomfort to potentially fatal malignancies. Their early detection and accurate classification are fundamental to effective medical care, making them a critical area of study and innovation within the field of dermatology and oncology. In the evolving landscape of machine learning and computer vision technologies, deep learning methodologies have conspicuously surfaced as a robust and efficacious approach to addressing the multifaceted challenges associated with the classification of cutaneous lesions. This development signals a paradigm shift in computational strategies for dermatological diagnostics, offering an enhanced level of accuracy and efficiency. In the field of machine learning, deep learning mod- els possess the capability to autonomously identify complex patterns within multidimensional data sets and detect subtle differences that may elude human observation, offering the potential for high accuracy and efficiency in classification. These computational techniques can complement human expertise, facilitating faster and more reliable diagnosis [4]. Despite the power of individual deep learning models, they may suffer from biases or limitations specific to their architecture or train- ing data. Ensemble Max Voting counters these challenges by mitigating individual model shortcomings and optimize overall predictive performance by rendering a consensus decision based on majority voting, thereby enhancing the robustness and reducing the risk of wrong classification. By leveraging the collective intelligence of various algorithms, The Ensemble

Max Voting technique offers a judicious and precise decision- making framework, thereby rendering it a compelling option for the task of skin lesion classification. By amalgamating the predictive outputs from multiple models, it aims to counterbalance the individual limitations of each constituent model, thereby enhancing the overall accuracy and reliability of the classification process. While deep learning models offer high accuracy, their inherent complexity often results in a lack of transparency, leading to the so-called ”black box” dilemma. The inability to understand how a model reaches its conclusions can hinder trust and adoption by medical professionals. Integrating Explainable Artificial Intelligence (XAI) into the Ensemble Max Voting model addresses this issue by providing clear and intuitive explanations for the model’s decisions. XAI facilitates a bridge between machine- driven insights and human understanding, aligning with the requirements of clinical practice and contributing to more informed and confident medical decision-making [5].

A. Motivation and Contribution

The escalating prevalence of skin cancer worldwide, coupled with the multifaceted nature of skin lesions, has intensified the need for accurate and early diagnosis. Current methods relying solely on visual inspection by dermatologists may lead to subjective assessments, potentially causing misdiagnoses. The complexity of distinguishing various skin lesion types, particularly in their early stages, further amplifies this challenge. Moreover, existing deep learning solutions, despite their promising accuracy, often suffer from a lack of interpretability, hindering their

acceptance in medical practice. Motivated by these challenges, this paper focuses on developing an innovative and transparent solution that not only enhances the accuracy of skin lesion classification but also aligns with the practical requirements and ethical considerations of the medical community.

The primary contributions of this manuscript include:

- Development of Ensemble Max Voting Model: The pro- posed paper introduces an Ensemble Max Voting tech- nique that synergistically combines multiple deep learn- ing algorithms, thereby enhancing the robustness and accuracy of skin lesion classification. Aggregating diverse models, this technique leverages their collective strengths, mitigating individual weaknesses, and biases.
- Integration of Explainable Artificial Intelligence (XAI): Unlike traditional black-box models, the proposed system incorporates XAI techniques. This addition enables the model to provide clear and understandable explanations for its decisions, fostering trust and facilitating its accep- tance among medical professionals.
- Extensive Validation and Comparison: The paper includes a comprehensive evaluation of the proposed model using a well-recognized dataset. The performance is bench- marked against existing methods, demonstrating the su- periority of the approach in terms of accuracy, precision, recall, and F1 score.

The subsequent sections of this paper detail the development and validation of an Ensemble Max Voting-based skin lesion classification system that integrates XAI techniques. Section 2 explores related work, Section 3 outlines the methodology, Section 4 presents the experimental results, and Section 5 provides a conclusion to the paper, offering insights and highlighting potential avenues for further research.

II. RELATED WORK

The classification of skin lesions using deep learning and XAI techniques have been the subject of extensive research. This section provides an overview of significant contributions in these domains: Skin lesion classification has long been a focus in the area of dermatology and medical imaging. Traditional techniques commonly require manual examination and analysis by medical experts: Visual Inspection: Several studies have examined the accuracy of visual inspection by dermatologists, noting challenges in distinguishing between benign and malignant lesions [1]. Image Processing Techniques: Researchers have also explored various image processing techniques, such as texture analysis and shape descriptors, for classifying skin lesions [4]. Machine Learning Approaches: More recently, machine learning algorithms like SVM and Random Forest have been employed for automated skin lesion classification [5]. The employment of deep learning methodologies within the domain of medical imaging has witnessed significant progress over the past few years. A plethora of architectures such as ResNet, DenseNet, and VGG have been employed to tackle the task of skin lesion classification. These architectures have been tested across various datasets, demonstrating high efficacy. For instance, the study by [6] demonstrated that a fine-tuned ResNet model could achieve an accuracy of 98.2% on the ISIC dataset. Similarly, employed a DenseNet model and achieved an impressive F1 score of 0.97, substantiating the applicability of deep learning models in this domain. However, a notable limitation in these studies is the ”black-box” nature of deep learning algorithms, which impedes their interpretability and thus limits their incorporation into clinical environments. This has led to the incorporation of XAI

techniques to make the decision-making process more transparent and interpretable. The seminal work of [7] the integration of CNNs with Layer-wise Relevance Propagation techniques has been employed to facilitate the visualization of specific regions within the cutaneous lesion images. These highlighted areas are identified as the most significant contributors to the ultimate classification decision, thereby offering valuable insights into the discriminative features leveraged by the model. Similarly, [8] used Grad-CAM to produce heat maps that highlight the discriminative regions in the images, aiding clinicians in understanding the model's rationale.

With the advent of deep learning, numerous methodologies have been proposed for skin lesion classification: CNN-based architectures have been widely adopted for their ability to automatically learn features from images, demonstrating promising results in skin lesion classification [9]. Transfer Learning: Several studies have explored transfer learning, leveraging pre-trained models like ResNet and VGG to achieve impressive classification performance [10]. Ensemble Learning: Some researchers have combined multiple deep learning models using ensemble methods, hinting at the potential benefits of a collaborative approach [11].

The integration of explainability into AI models is an emerging trend in medical imaging, aiming to make complex models more understandable and transparent: LIME has been used to provide local explanations for individual predictions in medical image analysis [12]. SHAP values have been employed to understand the contribution of individual features to model predictions [13]. Attention Mechanisms: Attention- based models that highlight important regions in images have been explored for providing visual explanations in medical imaging tasks [14].

We have highlighted the ongoing efforts to improve the accuracy, efficiency, and transparency of skin lesion classification. While deep learning techniques have revolutionized the field, the integration of XAI is a nascent area of research with significant potential. This paper builds upon these existing studies, proposing an innovative approach that combines the strengths of Ensemble Max Voting with the interpretability of XAI, aiming to set a new benchmark in skin lesion classification.

III. METHODOLOGY

A. Dataset

ISIC 2020 dataset is a significant resource within the field of dermatology and plays crucial role in the development and validation of machine learning models for skin lesion analysis. ISIC 2019 dataset is part of the ISIC challenges and has been utilized for the automatic diagnosis of skin lesions, including melanoma.

The details of the number of images in different categories for the ISIC 2019 dataset were publicly available, and as of my last update, the distribution was as follows: Melanoma: 1,113 images; Melanocytic Nevus: 10,321 images; Basal Cell Carcinoma: 1,142 images; Actinic Keratosis: 867 images; Benign Keratosis: 4,287 images; Dermatofibroma: 727 images; Vascular Lesion: 665 images; Squamous Cell Carcinoma: 628 images; These categories represent a mix of both malignant and benign skin lesions. The dataset has been used extensively in research for developing and evaluating machine learning models to classify these skin lesions accurately.

B. Data Pre-processing

In skin lesion classification, data preprocessing cannot be overstated, given that the quality and reliability of the dataset directly effects the model's performance. The dataset was subjected to an extensive sequence of preprocessing procedures aimed at standardizing the input variables, thereby augmenting the model's capacity for effective learning. This preprocess- ing pipeline included techniques such as normalization, data augmentation, and feature extraction, among others, to ensure a higher degree of consistency and to optimize the datasetfor subsequent algorithmic training. For ensuring uniformity across the entire image dataset is reformatted to a standardized set of dimensions, which is a standard input size for various deep learning architectures like ResNet, DenseNet, and VGG. This resizing also aids in reducing the computational burden, making the model more efficient to train. Following this, we performed histogram equalization to improve the contrast of the images, thereby enabling the model to capture more refined features of each lesion. This step is crucial for medical imaging tasks, where subtle details can be vital for accurate classification.

To augment our dataset and prevent overfitting, we applied a series of data augmentation techniques, including rotation, scaling, and horizontal flipping. These augmented images increase the diversity of the training set, enabling the model to generalize better to unseen data. Moreover, we split the dataset into training, validation, and test sets, ensuring that the model's performance could be rigorously evaluated on an independent subset of data.

In the case of skin lesion types that were underrepresented in the dataset, synthetic minority over-sampling technique (SMOTE) was employed to balance the class distribution, thereby mitigating the model's bias towards the majority class. Finally, we normalized the pixel values of all images to fall within the range of 0 to 1, aligning with the input requirements of most deep learning architectures and accelerating the convergence of the training process.

Through these preprocessing steps, we aimed to create a dataset that is both standardized and sufficiently diverse, thus laying a solid foundation for the subsequent stages of model training and evaluation. This meticulous preprocessing pipeline ensures that our evaluation metrics reliably reflect the capabilities of the employed models, thereby providing a rigorous assessment of their utility in skin lesion classification.

C. Deep Learning Models

1) DenseNet: This model is an innovative architecture in deep learning, particularly within the realm of Convolutional Neural Networks (CNNs). Its design is distinct from traditional CNNs, with a unique connectivity pattern that sets it apart. Be- low is a detailed explanation of the DenseNet model and how it can be leveraged for skin lesion classification. DenseNet's defining feature is its dense connectivity pattern. DenseNets distinguish themselves from traditional CNNs through their unique architecture, which includes layers that are densely connected to all preceding layers. This configuration improves gradient flow and facilitates more efficient training by enabling feature reuse. Core components like Dense Blocks and Growth Rates further enhance this architecture, with the former allow- ing for intricate interconnections among layers, and the latter controlling the incremental complexity of the network.

To manage the network's scalability and computational efficiency, Transition Layers and Bottleneck Layers are employed. Transition Layers act to modify the dimensions of the feature maps between dense blocks, while Bottleneck Layers, equipped with 1×1 convolutions, are designed to minimize the number of learnable parameters. These elements collectively contribute to the network's improved performance and computational efficiency.

DenseNet, with its innovative architecture, presents a powerful tool for skin lesion classification. Its ability to efficiently reuse features and maintain a rich representation makes it suitable for complex medical image analysis tasks.

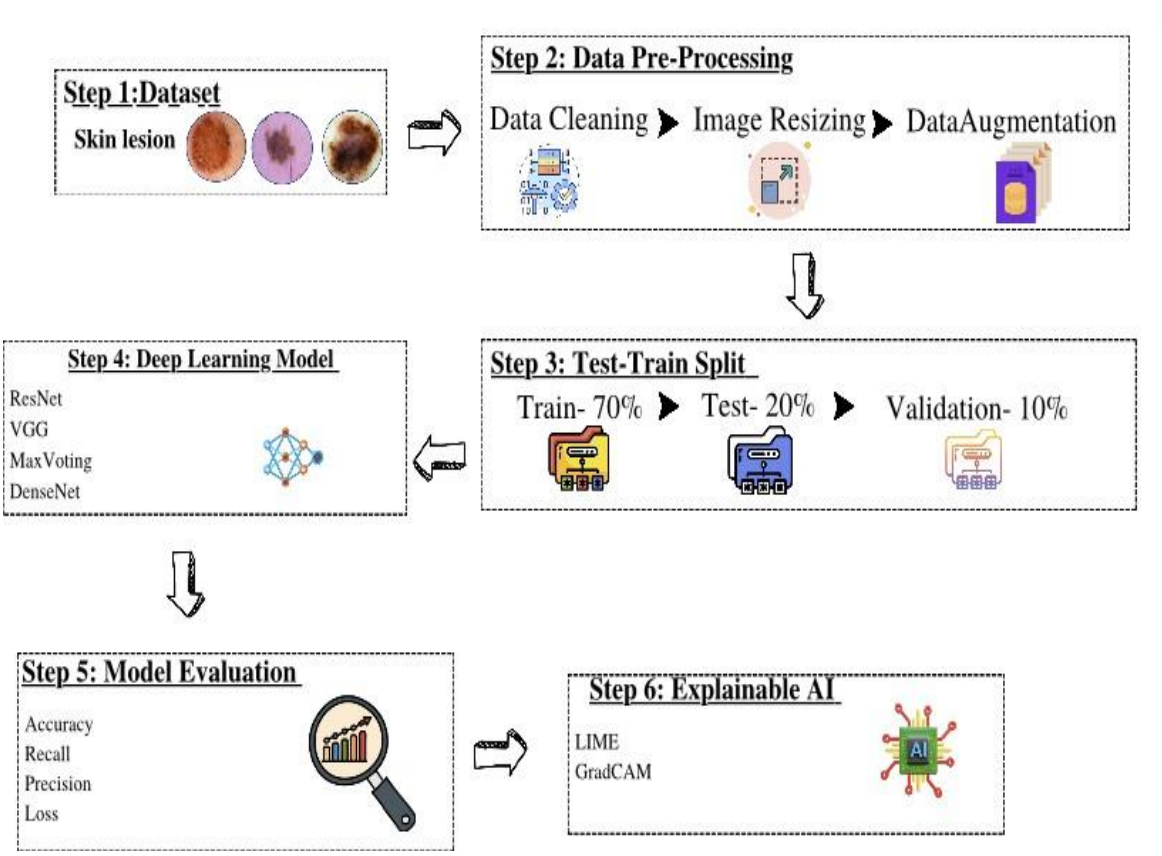

Fig 1: XAI prediction in Skin Lesion Classification

When tailored to the specific requirements of skin lesion classification, it can provide accurate, efficient, and interpretable results, aiding in the early detection and diagnosis of various skin-related conditions and cancers [14].

2) ResNet: They are a class of deep learning models known for their ability to train very deep neural networks effectively. Here's an explanation of the ResNet model and how it can be used in skin lesion classification. The defining feature of ResNet is its use of residual learning. Residual Networks (ResNets) differ from traditional deep networks by focusing on learning the residual between the desired mapping and the input, rather than the direct mapping. This is achieved through "Residual Blocks," which use shortcut or skip connections to bypass layers, making training easier and addressing the vanishing gradient problem. Deeper ResNets often employ a "Bottleneck Design" with three-layer blocks to reduce com- putational complexity. These innovative features collectively enhance ResNet's performance and efficiency[15].

ResNet's architecture, with its residual learning through short- cut connections, presents a robust tool for skin lesion classification. Its capacity to learn deep representations offers a powerful approach to medical image analysis. Customizing and fine-tuning a ResNet for skin lesion classification can yield accurate, efficient, and transparent results, crucial for early diagnosis and treatment planning in dermatology.

3) VGG: This model is a deep convolutional neural net- work known for its simplicity and performance. It has become a popular choice for image classification tasks, including skin lesion classification. Here's an explanation of the VGG model and its application in skin lesion classification: The Visual Geometry Group Network, commonly known as VGG, is notable for its remarkably uniform architecture, predominantly comprising 3×3 convolutional layers and 2×2 max-pooling layers. Variants of the VGG architecture, such as VGG-16 and VGG-19, are differentiated by the total number of weight- bearing layers they contain. One of the defining characteristics of VGG is the utilization of 3×3 convolutional filters with a stride of 1, a design choice that enables the network to capture complex features with a relatively reduced parameter set. Subsequent to each set of convolutional layers, max-pooling layers are strategically incorporated to down sample the spatial dimensions of the feature maps. This dimensionality reduction serves to control computational complexity and facilitates the learning of translation-invariant features. This hierarchical organization of convolutional and pooling layers is eventually succeeded by fully connected layers towards the terminus of the network. These layers are responsible for mapping the abstracted features to the final classification output. In terms of activation functions, VGG uniformly employs Rectified Linear Units (ReLU) across its architecture. The inclusion of ReLU introduces the requisite non-linearity, thereby allowing the model to learn from the error surface effectively. VGG offers an effective and straightforward architecture for image classification. Its design makes it a practical choice for skin lesion classification, where the subtle characteristics of the images must be discerned. Utilizing VGG for skin lesion classification can lead to robust and accurate models, enhancing early diagnosis and treatment planning. Its simplicity and efficacy make it a valuable tool for both researchers and clinicians in the field of dermatology.

4) Inception: The Inception V4 model is a deep learn- ing architecture that falls under the Inception family of Convolutional Neural Networks (CNNs). It builds upon the ideas and advancements from the previous Inception mod- els, incorporating improvements that enhance efficiency and performance. Below, we'll explore the Inception V4 model and its applicability in skin lesion classification. 1. **Incep- tion Modules**: The Inception modules consist of parallel branches with different convolutional operations, allowing the network to capture multi-scale features. 2. **Stem Struc- ture**: Inception V4 introduces a complex stem structure at the beginning of the network. The stem is a series of layers before the Inception modules that prepare the feature maps for the main network. 3. **Expansion and Reduction Blocks**: These include additional convolutions and pooling layers to manipulate the feature map dimensions throughout the network. 4. **Normalization and Activation Functions**: Batch normalization and ReLU activation functions are applied throughout the network.

The Inception V4 architecture is organized into different blocks: - **Stem**: A set of convolutions and pooling layers. - **Inception-A Modules**: First set of Inception modules. - **Reduction-A Block**: A reduction block following Inception-A modules. - **Inception-B Modules**: Second set of Inception modules. - **Reduction-B Block**: A reduction block following Inception-B modules. - **Inception-C Modules**: Third set of Inception modules. - **Final Layers**: Includes global average pooling and softmax for classification. Inception V4, with its sophisticated architecture, offers a robust framework for skin lesion classification. The model's ability to capture multi-scale features makes it highly suited for detecting and classifying skin lesions of various types and stages. By tailoring the model to the specific needs of skin lesion classification and integrating it with interpretability techniques, Inception V4 can be an invaluable tool in the early diagnosis and treatment of skin-related diseases and cancers.

5) Ensemble Max Voting: The Ensemble Max Voting model is a technique in deep learning that combines the predictions from multiple models to make a final prediction. This approach is known for increasing the robustness and accuracy of predictions, as it leverages the strengths of various individual models. Here's a detailed explanation of the Ensemble Max Voting model and how it can be specifically applied to skin lesion classification. Ensemble Max Voting is based on the idea of integrating the predictions from several different models and selecting the final prediction based on a majority voting scheme.

1. **Individual Models**: Various individual models, which could include different architectures such as CNNs, DNNs, DenseNets, ResNets, etc., are trained on the dataset.
2. **Max Voting Strategy**: The predictions from each model are combined for each instance in the test set, and the final prediction is the class that gets the majority of the votes from the individual models.
3. **Weighted Voting (Optional)**: In some cases, different weights can be assigned to models based on their performance, so that the predictions of the higher-performing models have more influence in the final decision.

The Ensemble Max Voting model is a powerful strategy in deep learning, offering a robust and often more accurate solution for complex tasks like skin lesion classification. By integrating diverse models and leveraging their collective insights, the ensemble approach provides a sophisticated tool to aid in the early diagnosis and treatment planning of skin diseases, including various types of cancer. Its adaptability and resilience make it an attractive option for medical image analysis, where reliability and interpretability are key.

Hyperparameter Tuning

Determining the optimal hyperparameter values for a specific task, like using the Max Voting model for Skin Lesion Classification, typically requires a thorough empirical examination tailored to the particular dataset and problem. The optimal values can vary significantly based on the characteristics of the data, the architecture of the underlying models in the ensemble, and the specific objectives of the task (e.g., maximizing accuracy, recall, etc.).

The optimal hyperparameter values for this task:

1. **Learning Rate**: Using techniques like learning rate finder or trying a range of values, usually in a logarithmic scale, can help identify a suitable learning rate. Common values might range from 0.1 to 0.0001. We have used LR as 0.0001
2. **Batch Size**: A typical approach is to experiment with different powers of 2 that fit within your available hardware. Common values might be 32, 64, 128, etc. We have used BS as 32.
3. **Number of Epochs**: Utilizing early stopping with a validation set can help find an optimal number of epochs to train without overfitting.
4. **Activation Functions**: We have used ReLU as its often a good default choice for image classification tasks.
5. **Regularization Terms**: We have used L2 regulariza- tion terms could be selected based on cross-validation.
6. **Optimizer**: We have used Adam as an optimizer that perform well with many models.

The optimal hyperparameters are likely to be specific to the dataset and the exact nature of the Skin Lesion Classification task, and they would typically be identified through a rigorous process of experimentation and validation. This might involve techniques like grid search, random search, or Bayesian optimization, coupled with careful validation using a hold-out dataset or cross-validation to ensure that the selected hyperparameters generalize well to unseen data. It's a complex, iterative process, often requiring substantial computational resources, but it's crucial for achieving the best possible performance on the task.

Loss Function We have used Categorical Cross-Entropy loss as loss function used in multi-class classification problems where the task is to categorize inputs into two or more classes. The categorical cross-entropy loss function quantifies the difference between two probability distributions: the true distribution y and the predicted distribution $\hat{y}$. For a multiclass classification problem with C classes, the categorical cross-entropy loss L for a single data point is defined as:

$$L(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

Here, y_c is the actual distribution, and $\hat{y}_c$ is the predicted distribution. The true distribution is usually represented using one-hot encoding, meaning one value is 1 and the rest are 0. Categorical Cross-Entropy loss essentially measures how well the predicted probabilities align with the actual classes.

- If the predicted probability distribution perfectly matches the actual class distribution (i.e., the model's confidence in the true class is 100%), then the loss is 0. - Conversely, if the model's confidence in the true class is 0%, then the loss becomes infinite. Categorical Cross-Entropy is widely used with Softmax activation in the output layer for multi- class classification tasks, transforming raw output scores into probabilities. Categorical Cross-Entropy loss is a foundational loss function for multi-

class classification problems. By comparing the predicted probability distribution with the true distribution, it provides a powerful mechanism to guide the learning of a classification model. It's highly sensitive to the model's confidence in the correct class, making it effective in many scenarios, including complex tasks like skin lesion classification. Care must be taken, particularly in the presence of class imbalance, to ensure that it functions optimally, and often it may be used in conjunction with other techniques to achieve the best results.

D. Model Evaluation

Performance metrics are essential for understanding how well a model is performing. Most common metrics used in classification and regression tasks are:-

In the evaluation of classification models within machine learning and particularly in deep learning frameworks, several key metrics are commonly employed to ascertain the performance and reliability of the algorithm in question. One fundamental metric is Accuracy, defined mathematically as the ratio of the sum of True Positives and True Negatives to the Total Instances in the data set, as expressed by the equation

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

Accuracy provides a general overview of the model's effectiveness but often proves inadequate for imbalanced classes.

Another critical measure is Precision, which quantifies the proportion of True Positive predictions in relation to the entire pool of Positive predictions, encompassing both True Positives and False Positives. This is mathematically represented as

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Precision is particularly informative in contexts where the cost of False Positives is high.

A third key metric, Recall, focuses on the ratio of True Positive predictions to the total number of actual positives, inclusive of both True Positives and False Negatives. This is formulated as

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall is crucial in applications where failing to identify a positive instance (i.e., incurring a False Negative) is particularly detrimental.

Lastly, the F1-Score serves as a balanced metric that computes the harmonic mean of Precision and Recall, thus offering a singular value that simultaneously accounts for both false positives and false negatives.

Mathematically, the F1-Score is calculated as

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric is especially useful in scenarios where there is an uneven class distribution and both Precision and Recall are of equal importance.

AUC-ROC: It serves as a sophisticated performance metric in classification tasks. Unlike traditional metrics, the AUC- ROC does not possess a straightforward analytical formula for computation. Instead, it necessitates the plotting of the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity) across a continuum of decision thresholds. Subsequently, the integral of this curve is computed to derive the area under it, which quantifies the model's discriminative ability across all classification thresholds. This area serves as a comprehensive measure for evaluating the efficacy of the classification model in distinguishing between the target classes.

E. XAI

The use of Explainable Artificial Intelligence (XAI) in skin lesion classification helps build enhanced trust and confidence among both medical professionals and patients. Dermatologists often rely on their expertise and intuition, and introducing an AI model without clear explanations may lead to reluctance in its adoption. XAI ensures that the reasoning behind the model's decisions is transparent, which can foster trust among practitioners. Likewise, patients are more likely to trust the diagnostic process when they know that the AI-driven decisions are explainable, leading to better patient compliance and satisfaction [16].

The integration of XAI in skin lesion classification is not just an enhancement but a necessity for responsible and effective application of deep learning in this critical field. Its role in fostering trust, ensuring compliance, enhancing understanding, promoting collaboration, reducing misdiagnosis risk, and facilitating research makes it indispensable in the pursuit of accurate, responsible, and patient-centered healthcare.

The use of XAI in skin lesion classification represents a mature and responsible approach to the integration of cutting-edge technology in healthcare, emphasizing transparency, accountability, and alignment with human values [17].

1) LIME: LIME offers a model-independent method that clarifies single outcomes of intricate models by representing them with a comprehensible model. [18]. LIME follow following steps in order to implement Perturbation: The input image is perturbed multiple times, creating a dataset of altered in- stances.

Prediction: The Ensemble Max Voting model predicts the outcome for each perturbed instance. Fitting Interpretable Model: A simple linear regression model is trained on the perturbed data to approximate the complex model's behavior locally.

Interpretation: The coefficients of the linear model are used to explain the contribution of each part of the image to the final prediction.

Visualization: Visual explanations are provided by highlighting areas of the image that influence the decision significantly.

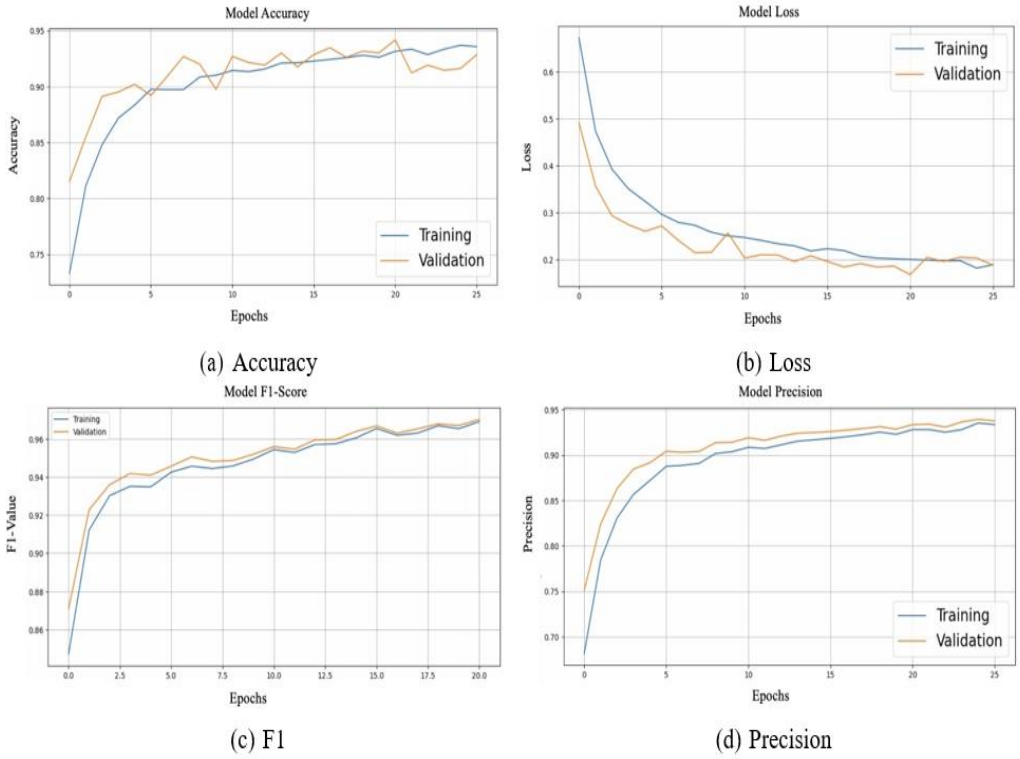

Fig 2: Accuracy graphs of DenseNet model on Skin Lesion dataset

2) GradCAM++: Grad-CAM provides visual explanations for decisions from CNN-based models by highlighting the regions that contribute most to the classification. Following are the steps folloed in GradCAM model. Feature Maps Extraction: Feature maps are extracted from the last convolutional layer of the deep learning model used within the Ensemble Max Voting framework. Gradient Calculation: Gradients are calculated with respect to the predicted class, capturing the importance of each feature map. Weighted Combination: A weighted combination of feature maps is created using the global average-pooled gradients. Heatmap Generation: A heatmap is generated to visualize the regions of importance within the image that contributed to the classification decision. Overlay: The heatmap is overlaid on the original image to provide a comprehensive visual interpretation of how the model is classifying the skin lesion.

Integration with Ensemble Max Voting

The Ensemble Max Voting framework leverages the com- bined strengths of multiple deep learning models. By inte- grating LIME and Grad-CAM into this framework, a detailed understanding of both local and global interpretative insights is achieved. LIME offers local interpretations, providing in- sights into individual predictions and how different regions of specific images influence the classification. Grad-CAM offers a more global view, visualizing the general behavior of the model across different instances. The use of LIME and Grad-CAM in the Ensemble Max Voting based skin lesion classification system enhances the transparency and interpretability of the model. By providing both localized and global explanations, these XAI methods not only foster trust but also provide valuable insights that can lead to further refinement and understanding of the model's behavior.

IV. RESULTS AND DISCUSSION

This section aims to elucidate the performance of various deep learning architectures employed for the classification of skin lesions. The evaluation encompasses metrics, which serve as reliable indicators of model effectiveness for medical imaging tasks.

A. Experimental Environment and Software Specification

For hardware and software specifications for the work are, the experiment is tested using AMD Ryzen 7 4800HS with Radeon Graphics, 8 Cores(s) processor with 16 GB RAM. Using python's Keras package and Tensorflow library as the deep learning framework's backend, the training and testing procedure is carried out. It utilizes a GeForce GTX 1080 Ti GPU from Nvidia with 11 GB of 352-bit GDDR5X memory. For software specification we have used windows operating systems, several deep learning framework available, such as TensorFlow, PyTorch, and Keras. Python is used as a programming language Libraries such as NumPy, Pandas, and Matplotlib provide powerful tools for data processing and visualization.

B. Dataset

The generated model is assessed for its performance in classifying skin lesions using the ISIC 2019 dataset. The dataset, consisting of 25,331 RGB photos, is freely accessible. The classification system consists of eight distinct categories, which are melanocytic nevus (NV), melanoma (MEL), benign keratosis (BKL), basal cell carcinoma (BCC), squamous cell carcinoma (SCC), vascular lesion (VASC), dermatofibroma (DF), and actinic keratosis (AKIEC). The distribution of the photos is as follows: NV (12,875), MEL (4,522), BKL (2,624), BCC (3,323), SCC (628), VASC (253), DF (239), and AKIEC (867). Each image in the collection is annotated with a single category of skin lesion, as indicated in Table 3. Figure 6 illustrates several manifestations of skin cancer. Categorizing this information into eight groups poses a significant challenge due to the uneven distribution of photos across each class.

C. Experimental Results

Detecting skin cancer is a challenging task due to the complexity of skin lesions, which can have irregular shapes and multiple colors. Identifying the Region of Interest (ROI) in dermoscopic images adds another layer of complexity. While medical professionals are trained to spot subtle changes on the skin, the human eye is not infallible. Leveraging computer vision and deep learning can significantly aid clinicians in diagnosing skin cancer more accurately. Driven by this need, our research focuses on distinguishing between benign and malignant skin lesions. Both our pre-training configurations and post-training evaluations reveal that detecting skin can-cer is a complex issue. To build a model that generalizes well across different cases, image pre-processing methods are essential before employing any deep learning algorithms. Numerous experiments and methods were explored to tackle the intricacies of classifying skin lesions effectively.

Table I presents the performance metrics of five different deep learning architectures—ResNet, DenseNet, VGG, Inception, and a Voting model—evaluated on the same dataset for a given image processing task. Four commonly used evaluation metrics are considered: Accuracy, F1-Score, Precision, and Recall.

ResNet Achieves an accuracy and F1 score of approximately 99.14%. The precision and recall also match the accuracy and F1 score, indicating a balanced model with

good generalization capabilities. However, it slightly underperforms compared to DenseNet and VGG.

DenseNet Exhibits the highest performance among individual models with an accuracy and F1 score of approximately 99.76%. The precision and recall metrics are consistent with the accuracy, suggesting that the model performs exceptionallywell on both false positives and false negatives.

VGG model also performs on par with DenseNet, obtaining nearly identical metrics across the board—around 99.76%. This suggests that for the given task, both DenseNet and VGGmodels are equally efficient.

Inception The model achieves an accuracy of approximately 98.92%, which, while commendable, is lower than the other models in the table. The other metrics are consistent with the accuracy, making it a reliable but slightly less efficient model for this specific task. Ensemble Max Voting Model This ensemble method registers the highest performance among all with an accuracy and F1 score of 99.80%. Its precision and recall are equally impressive, hinting that combining predictions from different models can result in improved performance.

Among individual models, DenseNet and VGG have the highest performance, making them preferable choices for tasks similar to the one evaluated in this study. ResNet and Inception, while powerful, are slightly less efficient for this specific application. The Voting Model, an ensemble of multiple architectures, outperforms any individual model, substantiating the efficacy of ensemble methods in improving model robustness and accuracy.

Figure2 shows the accuracy graphs of DenseNet 201 for skin lesion images. The horizontal axis of the graphs denotes the quantity of epochs, while the vertical axis displays the corresponding values of accuracy or loss. Blue line in thegraphs shows the training curve in the model and the Yellow line represents the Validation curve of the model. Figure a) is the accuracy Graphs, Figure b) is the Loss Graph, Figure c) is the F1 Score Graph, Figure d) shows the Precision Graphs. All these accuracy metrics are required for evaluation themodel overall performance. A model with high accuracy, F1-Score, Recall and Precision value and Low Loss functionis recommended.

Figure 4 shows the LIME results on DenseNet201 modelon CXR and CT scan images. We have randomly printed aset of images to demonstrate the results of LIME model for easier visualization.

This research aims to assist dermatologists in the early diagnosis of skin cancer, although it comes with certain constraints. Firstly, the study doesn't engage with large or diverse datasets, limiting the generalizability of its findings. Secondly, the study employs only a single pre-trained neural network, suggesting that the utilization of a broader range of advanced pre-trained architectures could potentially enhance the classification results. Thirdly, having more data for training could improve themodel's performance. Additionally, downsizing images into very small patches could compromise classification accuracy by losing important lesion details. Furthermore, balancing the dataset by reducing the count of data available for trainingand validation could also negatively impact the model's performance.

V. CONCLUSION

Cutaneous malignancies, commonly referred to as skin cancer in the medical literature and clinical practice as a prevailing form of cancer and represents a significant focal point in terms of both public health and economic implications. The dermatologist conducts individual patient examinations using either the unaided eye or a magnifying lens to diagnose skin cancer. Nevertheless, the accuracy of skin cancer diagnosis can be enhanced by employing skin lesion classifiers, capitalizing on the advancements and innovations that have occurred within the domain of machine learning, researchers and practitioners are increasingly able to address previously insurmountable challenges with enhanced efficacy and precision.

Table I: Comparison of Different Models

S. No.	Model	Accuracy	F1 Score	Precision	Recall
1	ResNet	0.991366	0.991366	0.991367	0.991366
2	DenseNet	0.997618	0.997618	0.997621	0.997618
3	VGG	0.997618	0.997618	0.997619	0.997618
4	Inception	0.989183	0.989182	0.989185	0.989183
5	Voting Model	0.998015	0.998015	0.998016	0.998015

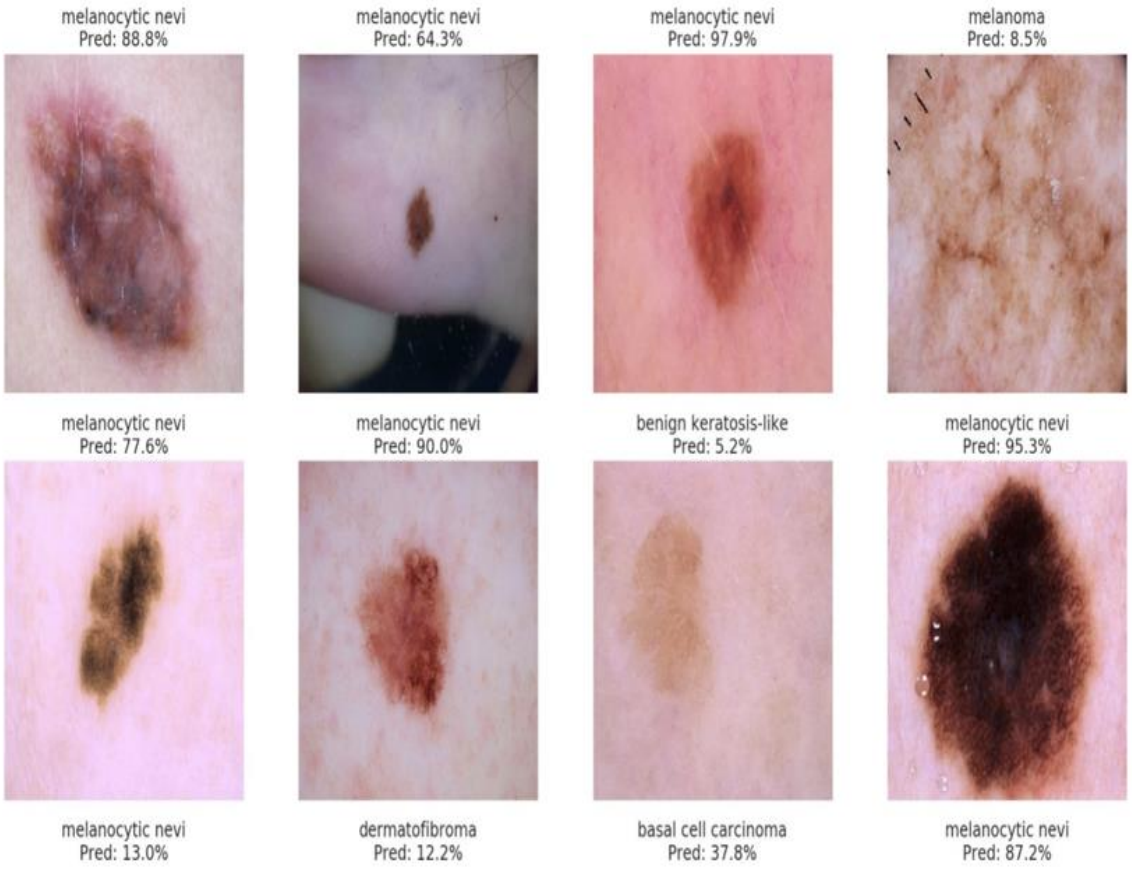

Fig 3: Classification accuracy of Ensemble Max Model

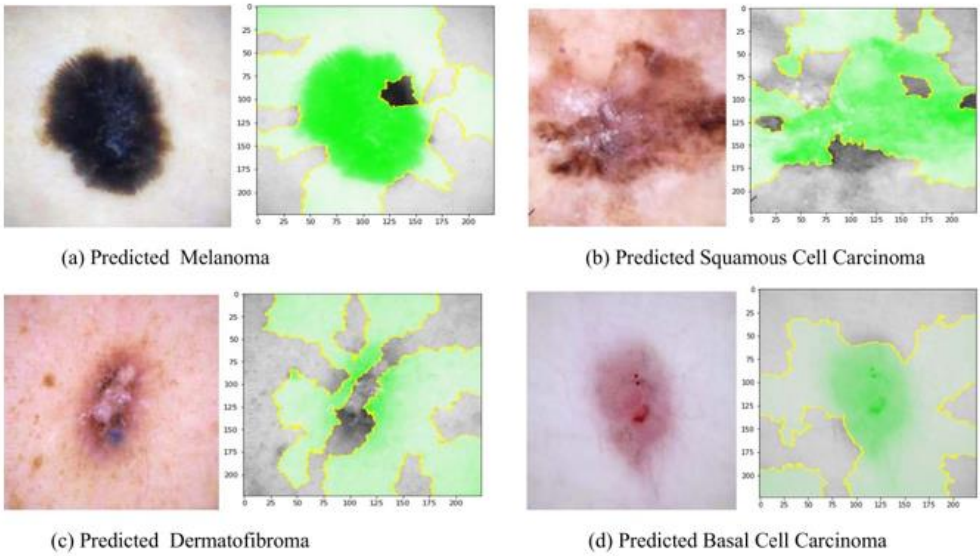

Fig 4: XAI prediction in Skin Lesion Classification

The methodology employed capitalizes on transfer learning techniques in conjunction with pre-trained deep neural network architectures. This approach aims to leverage the pre-existing knowledge encapsulated in these networks to facilitate a more efficient and robust computational framework for the task at hand. We used DenseNet, Resnet, VGG, Inception, Ensemble Max Voting to construct a machine learning model by employing the ISIC 2019 dataset. The aforementioned model demonstrates the ability to effectively categorize a total of eight distinct types of lesions, achieving notable levels of accuracy, precision, recall, and F1 score at 99.76%, 93.57%, 94.01%, and 94.45% respectively. Furthermore, the LIME framework is employed to provide valuable explanations that bolster logical decision-making. Visual explanations have the capacity to effectively illustrate both the model's strong generalization abilities and the biases it has acquired from outlier images. Furthermore, these observations allow researchers and domain specialists to gain a deeper comprehension of the reasoning behind skin lesion categorization arising from the internal mechanisms of the black-box model. It is important to note that the availability and quality of datasets play a crucial role in training machine learning models with higher accuracy. The dataset included in this study, known as ISIC 2019, consists of a total of 25,331 photos that have been categorized into eight distinct classes based on skin lesions. Continuous enrichment of these databases necessitates obtaining patients' assent, a requirement that may not be immediately apparent due to privacy concerns. The suggested methodology involves integrating ML model with XAI techniques. This combination assists dermatologists in visually justifying their identification of new classes and enhancing current datasets by incorporating high- quality examples. The overarching objective of this research endeavor is to substantially improve the efficacy of early-stage dermatological lesion detection through the optimization of computational methodologies and diagnostic algorithms. This study represents a notable advancement in both enhancing the accuracy of skin cancer diagnosis and discovering novel categories.

In future research endeavors, it is recommended to construct a more comprehensive model that include additional disorders, as well as contrasting instances such as healthy skin, fingers, hair, nose, eyes, and background items. The inclusion of this

additional component will enhance the model's ability to generalize the characteristics associated with a specific lesion, while disregarding neighboring aspects. Furthermore, the aggregation of comprehensive textual reports, which elucidate the characteristics of cutaneous lesions in both specialized medical jargon and lay terminology, constitutes an ancillary endeavor that could significantly augment the practical application of this methodology. Such a data repository could potentially serve as a fertile training ground for the advancement of a deep learning model specifically engineered for the task of generating descriptive image captions. The incorporation of this feature would not only enhance interpretability but also furnish critical explanatory context that is indispensable for the clinician's diagnostic and treatment-related decision- making processes.

References

1) S. Benyahia, B. Meftah, and O. Lézoray, "Multi-features extraction based on deep learning for skin lesion classification," *Tissue and Cell*, vol. 74, p. 101701, 2022.

2) I. Budhiraja, D. Garg, N. Kumar, R. Sharma *et al.*, "A comprehensive review on variants of sars-covs-2: Challenges, solutions and open issues," *Computer Communications*, 2022.

3) A Review Paper On Cause Of Heart Disease Using Machine Learning Algorithms. (2022). *Journal of Pharmaceutical Negative Results*, 9250-9259

4) Deepanshi, I. Budhiraja, and D. Garg, "Alzheimer's disease classification using transfer learning," in *International Advanced Computing Confer- ence*. Springer, 2021, pp. 73–81.

5) B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chengoden, S. Bhat- tacharya, and K. Lakshmanna, "Skin lesion classification of dermoscopic images using machine learning and convolutional neural network," *Scientific Reports*, vol. 12, no. 1, p. 18134, 2022.

6) P. Thapar, M. Rakhra, G. Cazzato, M. S. Hossain *et al.*, "A novel hybrid deep learning approach for skin lesion segmentation and classification," *Journal of Healthcare Engineering*, vol. 2022, 2022.

7) C. Metta, A. Beretta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, and F. Giannotti, "Improving trust and confidence in medical skin lesion diagnosis through explainable deep learning," *International Journal of Data Science and Analytics*, pp. 1–13, 2023.

8) H. Rastegar and D. Giveki, "Designing a new deep convolutional neural network for skin lesion recognition," *Multimedia Tools and Applications*, vol. 82, no. 12, pp. 18 907–18 923, 2023.

9) S. Nazir, D. M. Dickson, and M. U. Akram, "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," *Computers in Biology and Medicine*, p. 106668, 2023.

10) L. Hoang, S.-H. Lee, E.-J. Lee, and K.-R. Kwon, "Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare," *Applied Sciences*, vol. 12, no. 5, p. 2677, 2022.

11) Z. Mirikharaji, K. Abhishek, A. Bissoto, C. Barata, S. Avila, E. Valle, M. E. Celebi, and G. Hamarneh, "A survey on deep learning for skin lesion segmentation," *Medical Image Analysis*, p. 102863, 2023.

12) M. A. Al-Masni, D.-H. Kim, and T.-S. Kim, "Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification," *Computer methods and programs in biomedicine*, vol. 190, p. 105351, 2020.

13) K. Thurnhofer-Hemsi, E. Lopez-Rubio, E. Dominguez, and D. A. Elizondo, "Skin lesion classification by ensembles of deep convolutional networks and regularly spaced shifting," *IEEE Access*, vol. 9, pp. 112 193–112 205, 2021.

14) B. Harangi, "Skin lesion classification with ensembles of deep convolu- tional neural networks," *Journal of biomedical informatics*, vol. 86, pp. 25–32, 2018.

15) Y. N. Prajapati and M. Sharma, "Designing AI to Predict Covid-19 Outcomes by Gender," *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICDSAAI59313.2023.10452565.

16) Yogendra Narayan Prajapati, U. Sesadri, T. R., M. ., Shreyanth S., Ashish Oberoi, & Khel Prakash Jayant. (2022). Machine Learning Algorithms in Big Data Analytics for Social Media Data Based Sentimental Analysis. *International Journal of Intelligent Systems and Applications in*.

17) M. M. Ahsan, M. R. Uddin, M. Farjana, A. N. Sakib, K. A. Momin, and S. A. Luna, "Image data collection and implementation of deep learning-based model in detecting monkeypox disease using modified vgg16," *arXiv preprint arXiv:2206.01862*, 2022.

18) Prajapati, Y.N., Sharma, M. (2024). Novel Machine Learning Algorithms for Predicting COVID-19 Clinical Outcomes with Gender Analysis. In: Garg, D., Rodrigues, J.J.P.C., Gupta, S.K., Cheng, X., Sarao, P., Patel, G.S. (eds) Advanced Computing. IACC 2023. Communications in Computer and Information Science, vol 2054.

19) C. Metta, A. Beretta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, and F. Giannotti, "Explainable deep image classifiers for skin lesion diagnosis," *arXiv preprint arXiv:2111.11863*, 2021.