IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 March 2024, accepted 25 March 2024, date of publication 4 April 2024, date of current version 16 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3385099

RESEARCH ARTICLE

Attention to Monkeypox: An Interpretable Monkeypox Detection Technique Using Attention Mechanism

AVI DEB RAHA[1], (Graduate Student Member, IEEE), MRITYUNJOY GAIN[1], RAMESWAR DEBNATH[1], (Senior Member, IEEE), APURBA ADHIKARY[2], (Student Member, IEEE), YU QIAO[3], (Graduate Student Member, IEEE), MD. MEHEDI HASSAN[1], (Member, IEEE), ANUPAM KUMAR BAIRAGI[1], (Senior Member, IEEE), AND SHEIKH MOHAMMED SHARIFUL ISLAM[4]

[1]Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh
[2]Department of Information and Communication Engineering, Noakhali Science and Technology University, Noakhali 3814, Bangladesh
[3]Department of Artificial Intelligence, Kyung Hee University, Yongin 17104, Republic of Korea
[4]Institute for Physical Activity and Nutrition, Deakin University, Melbourne, VIC 3125, Australia

Corresponding authors: Anupam Kumar Bairagi (anupam@cse.ku.ac.bd) and Sheikh Mohammed Shariful Islam (shariful.islam@deakin.edu.au)

ABSTRACT In the wake of COVID-19, rising monkeypox cases pose a potential pandemic threat. While less severe than COVID-19, its increasing spread underscores the urgency of early detection and isolation to control the disease. The main difficulty in diagnosing monkeypox arises from its prolonged diagnostic process and symptoms that are similar to those of other skin diseases, making early detection and isolation challenging. To address this, the deployment of deep learning models on edge devices presents a viable solution for the rapid and accurate detection of monkeypox. However, the resource constraints of edge devices require the use of lightweight deep learning models. The limitation of these models often involves a trade-off with accuracy, which is unacceptable in the context of medical diagnostics. Therefore, the development of optimized deep learning models that are both resource-efficient for edge computing and highly accurate becomes imperative. To this end, an attention-based MobileNetV2 model for monkeypox detection, capitalizing on the inherent lightweight design of MobileNetV2 for effective deployment on edge devices, is proposed. This model, enhanced with both spatial and channel attention mechanisms, is tailored for rapid and early-stage diagnosis of monkeypox with better accuracy. We significantly improved the Monkeypox Skin Images Dataset (MSID) by incorporating a broader range of classes for similar skin diseases, thereby substantially enriching and diversifying the training dataset. This helps better distinguish monkeypox from other similar skin diseases, particularly in its early stages or when a detailed medical examination is unavailable. To ensure transparency and interpretability, we incorporated Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME) to provide clear insights into the model's diagnostic reasoning. Finally, to comprehensively assess the performance of our model, we employed a range of evaluation metrics, including Cohen's Kappa, Matthews Correlation Coefficient, and Youden's J Index, alongside traditional measures like accuracy, F1-score, precision, recall, sensitivity, and specificity. The attention-based MobileNetV2 model demonstrated impressive results, outperforming the baseline models by achieving 92.28% accuracy in the extended MSID dataset, 98.19% in the original MSID dataset, and 93.33% in the Monkeypox Skin Lesion Dataset (MSLD) dataset.

INDEX TERMS Attention, channel attention, Monkeypox, skin disease classification, spatial attention, MobileNetv2, ResNet, VGG.

The associate editor coordinating the review of this manuscript and approving it for publication was Yu Liu.

I. INTRODUCTION

In the current global health scenario, as the world continues to emerge from the shadow of the COVID-19 pandemic, there is growing concern regarding a new challenge: the outbreak of monkeypox. Monkeypox arises from an infection with the monkeypox Virus, a pathogen commonly found in monkeys [1]. Discovered in 1958 and re-emerging in the Republic of Congo in 2014, monkeypox, though less known than Ebola or Zika, could escalate into a significant global health concern [2]. The World Health Organization (WHO) has raised the alarm over the monkeypox outbreak, classifying it as a global health emergency [3]. Its spread is ongoing, with an increasing number of cases reported daily. Between January 1, 2022, and August 9, 2023, the WHO reported a total of 89,308 laboratory-confirmed monkeypox cases and 152 deaths across 113 countries in all six WHO regions [4]. During this period, a noticeable 1.2% increase in cases (1,020 new cases) and three new deaths were reported. In the fortnight leading up to August 9, 2023, there was an increase in reported cases in the Western Pacific, European, and American regions [4].

The clinical diagnosis of monkeypox presents a significant challenge due to its dermatological manifestations, which closely resemble those of several other conditions. The disease typically begins with a rash, progressing to pustules and scabs. This rash is initially most prominent on the face before spreading to other body parts. This presentation is similar to chickenpox, another condition known for its itchy, vesicular rash that transforms into fluid-filled lesions and then scabs. However, chickenpox rashes are generally more widespread across the entire body. Additionally, the early stages of monkeypox can be confused with measles, a contagious disease characterized by a red, blotchy rash. Like monkeypox, the measles rash starts on the face and then spreads downwards. Non-infectious skin conditions such as eczema and lupus further complicate the diagnostic process. Eczema, with its inflamed, itchy patches of skin, could be mistaken for the early stages of monkeypox. Lupus, known for its diverse range of rashes — including the distinctive butterfly-shaped facial rash or disc-like lesions — can also be confused with monkeypox during certain stages of its presentation. Other diseases like molluscum contagiosum, which manifests as tiny, firm bumps with a central indentation, and scabies, characterized by intense itching and a pimple-like rash from mite infestation, add to the diagnostic confusion due to the similarity in their skin lesions to those of monkeypox. Given these similarities, accurately diagnosing monkeypox can be particularly challenging for non-professionals. This underscores the importance of medical consultation and expert evaluation for accurate diagnosis, especially in these overlapping dermatological features.

The primary challenge in diagnosing monkeypox lies in its prolonged diagnostic timeline and the non-specific nature of its symptoms, which often resemble those of other skin diseases. This difficulty is further exacerbated by the limited availability of polymerase chain reaction (PCR) testing, a crucial component for the rapid diagnosis and containment of the disease's spread [1], [3]. Despite the relatively low mortality rate of monkeypox, which ranges from 1–10% [3], early detection remains critical. It plays a pivotal role in preventing further transmission, managing outbreaks, and implementing effective measures such as isolation and treatment [5]. In this context, deep learning emerges as a promising alternative, offering a viable solution to overcome these challenges. By leveraging the advanced pattern recognition capabilities inherent in deep learning algorithms, it becomes possible to swiftly identify and differentiate the visual characteristics of monkeypox from other similar skin conditions through image analysis. This approach can significantly reduce the diagnostic timeline, providing rapid preliminary diagnoses that could lessen the reliance on PCR testing for initial screenings. Furthermore, deep learning models, trained on a diverse dataset of skin disease images, including those specific to monkeypox, enhance their accuracy and robustness in diagnosing across a wide range of real-world scenarios.

Recently, deep learning (DL) techniques [6] have gained recognition as an effective method for image analysis and pattern recognition [7], proving particularly useful in detecting various diseases. As a branch of machine learning, DL utilizes multiple layers of artificial neural networks (ANNs) to extract features from images and make predictive analyses [8]. Convolutional Neural Networks (CNNs), a DL algorithm, have been effectively employed in numerous medical imaging applications. These include the classification of skin lesions, detection of breast cancer, and identification of lung nodules [9]. Automated systems leveraging machine learning (ML) and deep learning to address these challenges offer promising solutions. Various convolutional neural networks (CNNs) have proven effective in differentiating images of various diseases. This advancement implies that conditions like thyroid [10] cancer, pneumonia [11], and COVID-19 [12] can now be identified autonomously, without the direct involvement of a doctor [13].

In the context of infectious diseases like monkeypox, the application of DL becomes even more significant [1], [14]. The ability of CNNs to discern patterns and features in medical images offers a promising solution to the challenges posed by diseases that manifest with visual symptoms. These automated systems, based on ML and DL, have the potential to transform the diagnostic process by providing accurate and efficient disease identification. The effectiveness of CNNs in differentiating images of diverse medical conditions suggests a promising future for autonomous medical diagnostics. This emerging trend underscores the potential of deep learning in tackling public health challenges, particularly in detecting emerging infectious diseases like monkeypox. By harnessing the capabilities of advanced CNNs, it is possible to develop robust diagnostic tools that enhance the accuracy of disease

detection and contribute to faster and more efficient disease management. However, for efficient early detection of infectious diseases like monkeypox, deploying trained deep learning models on edge devices (e.g., mobile phones and tablets, etc,.) is essential. As this strategy offers several key advantages: it enhances accessibility, enabling its widespread application across various regions, including those that are remote; it allows for rapid response capabilities by providing immediate feedback on potential health concerns; and it delivers significant relief to healthcare systems through the automation of the initial stages of disease detection. Crucially, by implementing these models directly on edge devices, the privacy of patient data is robustly protected. Local processing ensures that sensitive information is not transmitted to external servers, maintaining confidentiality and security. With DL integrated into everyday devices like smartphones and tablets, patients gain tools for early disease detection [15]. This patient-centric approach is instrumental for diseases like monkeypox, where early identification and self-quarantine are crucial to controlling spread. These edge devices allow for convenient health assessments, promoting early self-diagnosis and encouraging timely medical consultation, thus reducing exposure risk and relieving healthcare systems. However, employing deep learning models in edge devices for medical diagnostics, especially for infectious diseases like monkeypox, presents a unique set of challenges. The primary obstacle lies in the limited computational resources and memory capacity of these devices, which are insufficient for running sophisticated DL models like VGG-16, VGG-19 or ResNet-152. These models, while effective, demand substantial processing power and memory—requirements that exceed the capabilities of most edge devices [16], [17]. Moreover, the challenge is compounded by the diversity of edge devices, each with varying hardware capabilities. This heterogeneity necessitates the development of adaptable and scalable lightweight deep learning models. However, a further obstacle emerges when these lightweight models are tasked with the early detection of monkeypox, as distinguishing it from diseases with similar symptoms is inherently challenging. The early signs of monkeypox often mirror those of other infections, such as chickenpox, measles or molluscum-contagiosum, making accurate diagnosis critical yet difficult to achieve. This necessitates not only the development of models that are compact and capable of running on various devices but also ones that are precise in their diagnostic abilities. However, traditional lightweight models (i,e,. ShuffleNet, MobileNet, GoogLeNet etc.) typically operate on simpler algorithms that prioritize efficiency over accuracy. While this trade-off is acceptable in applications with less critical outcomes, it becomes a significant limitation in medical diagnostics, where the stakes include human health and the potential for widespread disease transmission. Thus, there's a pressing need for the development of models that not only offer the high accuracy required for reliable disease diagnosis but also maintain the efficiency necessary for deployment on edge devices, ensuring they are both effective in diagnosing complex diseases like monkeypox and optimized for the computational constraints of edge computing.

Building on this necessity, our study introduces an attention-based lightweight deep learning model tailored explicitly for detecting and classifying various skin diseases, including monkeypox. This lightweight model ensures high efficiency and broad deployability across diverse platforms, including mobile devices. This aligns seamlessly with trends in integrating DL into patient-operated devices, facilitating widespread accessibility for early detection and timely intervention. The contributions of this study are summarized as follows:

- To significantly enhance the robustness of monkeypox detection through image classification and to substantially enlarge the training dataset, we meticulously extended the MSID dataset. This enriched dataset is now more harmonized, adeptly accommodating the diverse array of data nuances such as labeling, lighting, image quality, size, and resolution from various sources. The Extended MSID (EMSID) dataset now encompasses eight classes, thus elevating the intricacy and precision of the detection process.
- We proposed an attention mechanism-based deep learning model to detect and classify monkeypox disease more accurately. We utilized the MobileNetV2 architecture as the backbone for our model, enhancing it with both spatial and channel attention mechanisms. These attention modules allow the model to focus on the most relevant features within the images, significantly improving its ability to discern between monkeypox and other skin conditions. Applying these attention mechanisms within the MobileNetV2 framework creates a robust yet lightweight model ideal for deployment in resource-constrained environments and edge devices.
- Through a rigorous series of tests and analyses, we determined the most effective configurations for applying spatial and channel attention in MobilenetV2 Network to detect monkeypox.
- To validate the efficacy of our approach, we conducted extensive experiments comparing our proposed model with several state-of-the-art deep learning models, including ResNet-152, VGG-19, GoogLeNet, AlexNet, ShuffleNetV2 and the standard MobileNetV2. The performance of each model was evaluated using a comprehensive set of metrics: accuracy, precision, recall, F1-score, specificity, sensitivity, Matthews Correlation Coefficient (MCC), and Youden's J Index. These metrics provided a holistic view of the models' capabilities in accurately classifying monkeypox images.
- To ensure the interpretability of the deep learning model's decision-making process, we used Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations

(LIME). These techniques allowed us to visualize the areas within the images most influential in the model's classifications, thereby providing valuable insights into the reasoning behind the model's predictions.

- The results of our experiments demonstrated that our attention-based MobileNetV2 model outperformed the baseline models regarding accuracy, precision, and other vital metrics. This superior performance highlights the potential of our model in real-world applications, particularly in the early detection of monkeypox, which is crucial for timely treatment and containment of the disease.

The other parts of this study are organized as follows: The related works are presented in Section II. The methodology of the proposed model is presented in Section III. An operational workflow has been described in section IV. Baseline methods and performance matrices are discussed in Section V. The experimental results and performance comparisons are analyzed and discussed with its explanation in Section V-B. Finally, the conclusion and future course of this work is presented in Section VII.

II. LITERATURE REVIEW

Recent studies have increasingly employed deep learning techniques to analyze and identify monkeypoxes from image data. A variety of deep learning architectures, including ResNet-50 [18], VGG-19 [19], InceptionV3 [20], and MobileNetV2 [21], have been utilized in these investigations. For instance, in one study [14], three deep learning models - VGG-16, ResNet-50, and InceptionV3 - were adapted and trained using the Monkeypox Skin Lesion Database (MSLD). Due to limited data availability, data augmentation techniques were applied, leading to varied accuracy results: ResNet-50 achieved the highest accuracy at 82.96 (±4.57%), followed by VGG-16 with an accuracy of 81.48 (±6.87%), and InceptionV3 showed lowest accuracy at 79.26 (±1.05%). An advanced method for classifying monkeypox skin lesions using transfer learning has been developed in the work of [22]. They focused on providing a solution for areas with limited PCR testing by using Deep Learning to automate skin lesion detection. However, they only used three classes for their study, which limits the model's ability to differentiate monkeypox from a wider array of skin conditions. This raises concerns about the model's generalizability and effectiveness in real-world scenarios, where a diverse range of skin lesions might be encountered. Also, the model lacks interpretability. In their research, Yasmin et al. [5] developed 'PoxNet22', a machine learning-based model for diagnosing monkeypox with image analysis techniques. Utilizing transfer learning and data augmentation, they trained several deep learning models and found 'PoxNet22' to be the most effective, achieving 100% precision, recall, and accuracy. This innovation marks a significant advance in precise and reliable disease detection. However, this model also lacks explainability. Additionally, they have implemented a binary classification system, which, while effective in distinguishing monkeypox from non-monkeypox cases, limits the detection process's ability to differentiate between various skin conditions. This binary approach could lead to less nuanced decision-making when facing a broader dermatological disease spectrum. The authors of [23] employed Local Interpretable Model-Agnostic Explanations (LIME) to verify and interpret the performance of various deep learning (DL) models. These models included VGG-16, InceptionNetV2, ResNet-50, ResNet-101, MobileNetV2, and VGG-19 and were evaluated using generalization and regularization techniques. The objective was to enhance the accuracy and reliability of these models for diagnosing monkeypox. However, the study was again limited to only two classes, which might impact the model's capability to handle more complex, real-world scenarios.

The authors of [24] introduced a novel deep feature engineering architecture comprising multiple stages, including nested patch division, deep feature extraction, and feature selection using various analytical methods. The classification was conducted using an SVM classifier with 10-fold cross-validation, and results were consolidated through iterative hard majority voting (IHMV) and optimized by a greedy algorithm.The proposed model demonstrated a high classification accuracy of 91.87% on the dataset, marking the highest performance among 70 tested outcomes. This achievement underscores the potential of the presented framework in enhancing the detection and management of monkeypox through advanced machine learning techniques.

In a recent study [3], researchers focused on developing and validating deep learning models for the early detection of monkeypox, a critical step in controlling its spread. The study tested five standard pre-trained deep learning models: VGG-19, VGG-16, ResNet-50, MobileNetV2, and EfficientNetB3. Performance metrics such as accuracy, recall, precision, and F1-score were employed to assess the models' efficacy. Results revealed that the MobileNetV2 model demonstrated superior performance, achieving a notable accuracy of 98.16%, along with high recall, precision, and F1-score. Furthermore, validation using different datasets affirmed the model's consistency, with MobileNetV2 showing the highest accuracy. These findings highlight the potential of machine learning for early and accurate detection of monkeypox, positioning MobileNetV2 as a highly effective tool for clinical diagnostics. The study's comprehensive approach to model evaluation and validation contributes significantly to the growing body of research in medical image analysis, particularly in infectious disease diagnosis. In their work, Bala et al. [1] developed a specialized monkeypox research dataset named the Monkeypox Skin Images Dataset (MSID). This dataset is segmented into four distinct classes: ''Monkeypox,'' ''Chickenpox,'' ''Measles,'' and ''Normal,'' providing a comprehensive range for analysis. Building on this, they introduced a tailored version of the DenseNet-201 CNN architecture, aptly named 'MonkeyNet'. This model underwent rigorous testing on the original and augmented versions of the MSID, demonstrating its ability to accurately

diagnose monkeypox with an impressive 93.19% and 98.91% accuracy, respectively. A notable feature of 'MonkeyNet' is its integration of Grad-CAM visualizations, highlighting the model's decision-making areas in each image. This aspect not only confirms the model's diagnostic precision but also offers valuable insights for clinicians in identifying infected regions, enhancing the utility of the model in clinical settings. In [25], a Generalization and Regularization-based Transfer Learning (GRA-TLA), was specifically designed for binary and multiclass classification of monkeypox through image analysis. The authors rigorously tested this approach on ten different convolutional neural network (CNN) models across three separate studies. The findings revealed that the combination of GRA-TLA with the Extreme Inception (Xception) model yielded an accuracy between 77% to 88% for binary classification, identifying individuals with or without monkeypox. For multiclass classification, which is inherently more complex, the Residual Network (ResNet)-101 model stood out, achieving an impressive accuracy range of 84% to 99%. The work of [26] used ViT to distinguish between monkeypox and chickenpox. The ViT model achieved 93% accuracy, precision, recall, and F1-score in testing, showcasing the effectiveness of advanced deep learning models in medical imaging. The combination of transfer learning and image augmentation not only improved detection of monkeypox and chickenpox but also mitigated data scarcity issues, surpassing previous studies and CNN models in accuracy. In [27], the authors investigate the use of Deep Learning for detecting monkeypox (mpox) from smartphone images, responding to its classification as a global health emergency by the WHO. They utilize Transfer Learning to address the lack of extensive mpox image datasets, creating a refined dataset for analysis. Through rigorous evaluation of Convolutional Neural Networks (CNNs), MobileNetV3Large emerged as the top model, achieving an F-1 score of 0.928 in binary classification and 0.879 in multi-class tasks. Post-quantization optimizations reduced the model size by over two-thirds and decreased inference time from 0.016 to 0.014 seconds, with only a minimal reduction in the F-1 score of 0.004.

Most previous studies on monkeypox detection using deep learning models have primarily focused on limited classes and supplemented them with data augmentation techniques to artificially expand the training data. Employing a limited number of classes in previous studies on monkeypox detection using deep learning techniques introduces several challenges and limitations. Primarily, a model trained and tested on datasets with only a few classes may not adequately capture the complexity and variability inherent in real-world scenarios.

Furthermore, the reliance on data augmentation in many of these studies can be a double-edged sword. While data augmentation techniques like rotation, scaling, and flipping can help increase the training dataset's size and introduce variability, they may also inadvertently introduce artifacts or distortions that are not representative of actual clinical conditions. This can result in models overly optimized for the augmented data and may not perform as well when presented with authentic, unmodified images.

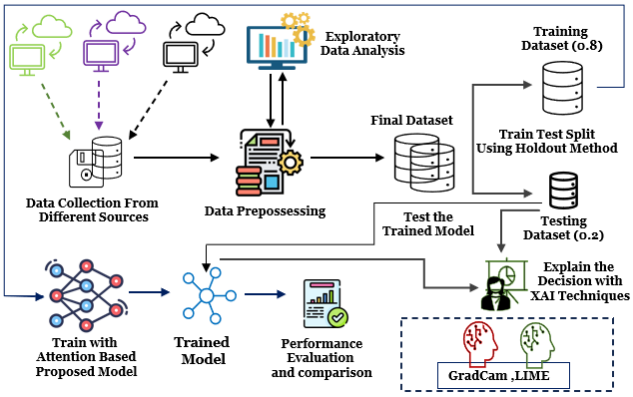

FIGURE 1. Overview of the proposed deep learning-based diagnostic framework for monkeypox detection.

In contrast, our approach to developing a model for monkeypox detection has focused on creating a more robust and generalizable system. By not relying on data augmentation, we ensure that our model is trained on authentic, unaltered images, which better represent clinical cases' true diversity and complexity. This approach enhances the model's ability to generalize to new data, making it more reliable and effective in practical settings. Moreover, the performance of our model in both the EMSID and MSID datasets, achieving accuracy rates of 92.28% and 98.19%, respectively, further underscores its effectiveness. These results are particularly noteworthy given that they were obtained without data augmentation.

III. METHODOLOGY

This section provides a detailed overview of the research methodology followed in this study. The process begins with acquiring data from various sources, an essential step for building a robust dataset representative of the problem scope. This data is then subjected to prepossessing, as illustrated in Fig. 1, to ensure data quality and remove any inconsistencies or errors. Once the prepossessing stage is completed, we obtain the finalized dataset. This dataset is then divided into two subsets— a training dataset and a testing dataset, following standard machine learning procedures.

This step is pivotal for ensuring the model's transparency, accountability, and interpretability. By employing XAI techniques, we aim to make the model's decisions understandable to technical experts and non-experts alike, thereby enhancing trust and facilitating broader adoption of the model in real-world applications, especially in the medical field.

A. DATASET COLLECTION

In this research, our dataset was primarily sourced from two premier databases: the Monkeypox Skin Images Dataset (MSID) [1] and DermNet [28]. The MSID features four

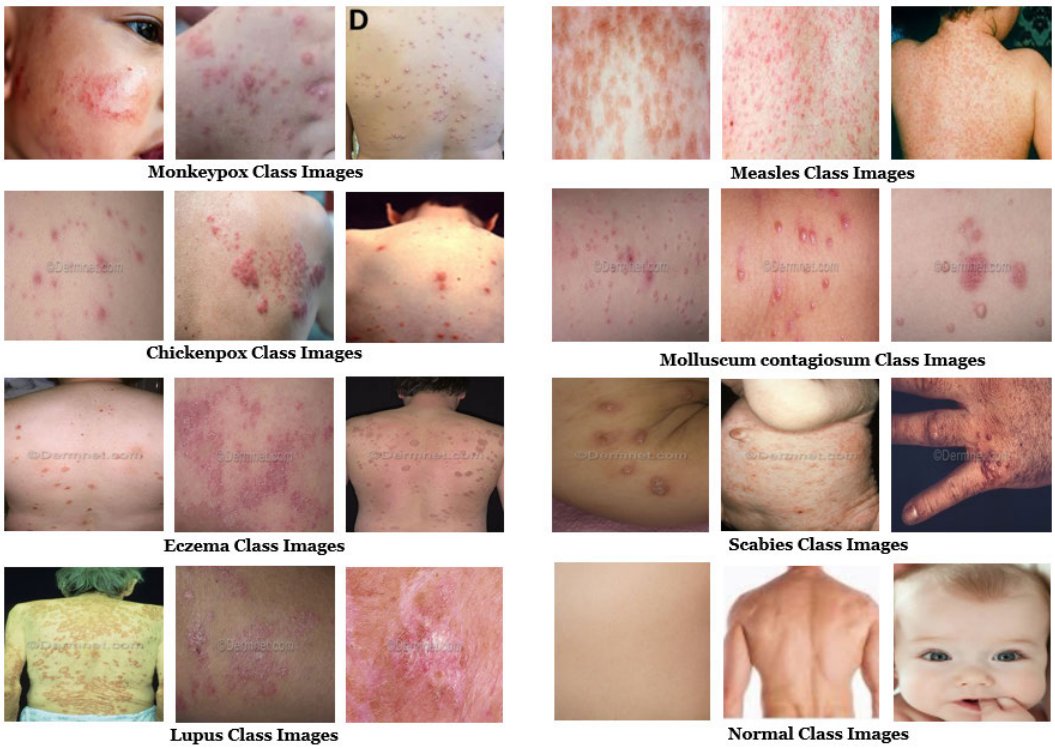

FIGURE 2. Instances of the EMSID.

categories of images: monkeypox, chickenpox, measles, and normal skin conditions. The images in this dataset were carefully curated from trustworthy online sources, including recognized health websites, newspapers, and scholarly journals. While the MSID dataset already presents a complex set of skin conditions, it is worth noting that some diseases may appear similar to the naked eye. To further diversify the range of skin diseases and increase the dataset's complexity, we also included images from Dermnet's online public repository. The final merged dataset is comprised of eight classes:

- **Monkeypox**: Presents with a rash that develops into pustules and then forms a scab, which later falls off. It often appears first on the face and then spreads.
- **Chickenpox**: Characterized by itchy red spots that turn into fluid-filled blisters. They eventually scab over and are usually seen all over the body. Like monkeypox, chickenpox causes a vesicular rash.
- **Measles**: Manifests as a red, blotchy rash that usually starts on the face and spreads downward. The red, blotchy rash can be easily confused with monkeypox.
- **Eczema**: Appears as patches of inflamed, itchy skin, which may be red in lighter skin tones or darker in darker skin tones. It can occur anywhere but is often found on arms and behind the knees. While generally easy to distinguish from monkeypox by a medical professional, to the untrained eye, any rash might be a cause for concern.
- **Lupus**: Rash can be very diverse, including a butterfly-shaped rash on the face or disc-like lesions elsewhere on the body. The pimple-like rash could potentially be confused with monkeypox.
- **Molluscum**: Presents as small, round, and flesh-colored bumps that are usually smooth and firm. They often have a small indentation in the center. Both the molluscum and monkeypox produce skin lesions that could appear similar at certain stages.
- **Scabies**: Characterized by intense itching and a pimple-like rash. The condition is caused by mites burrowing into the skin.
- **Normal**: The normal skin class contains no skin diseases in the skin. This class helps to distinguish healthy skin from pathological conditions.

While they may have similarities like redness, bumps, or rash, each condition has unique characteristics that trained healthcare professionals or a trained neural network model can identify. However, identifying them, especially to the untrained eye, is challenging, leading to potential misdiagnoses. Some image instances of the collected dataset have been given in Fig. 2. The figure shows that it is tough for untrained people to distinguish between skin diseases. In our collected dataset, there are a total of 1, 285 instances. The individual instances and their ratio for each class are shown in Fig. 3. Fig. 3 shows that the most prevalent classes in our dataset are ''Monkeypox'' and ''Normal'', comprising approximately 21.7% and 22.8% of the total instances, respectively. This is followed by ''Lupus'' and ''Chickenpox'', each contributing to around 12% of the total dataset. The other classes, namely ''Eczema'', ''Scabies'',

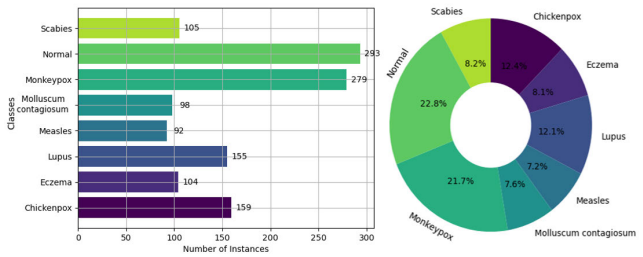

FIGURE 3. Data distribution in the EMSID.

''Measles'', and ''Molluscum contagiosum'', have relatively fewer instances, roughly 8% to 12% individually. It should be noted that the presence of a ''Normal'' skin class is essential for the model to correctly differentiate between pathological and non-pathological conditions correctly, thereby reducing the chance of false positives. The color-coded bar chart (Fig. 3) serves as a visual summary of the distribution of instances across different classes. The colors in the bar chart are consistent with those in the pie chart, providing a cohesive visual representation for easier interpretation.

B. DATA PREPROCESSING

Before model training, several preprocessing steps were performed on the dataset to ensure that the data was suitable for effective learning. Notably, no data augmentation techniques were employed in this study to preserve the genuine nature of the medical images. This approach was chosen to ensure that the model learns from authentic and real-world examples, which is crucial for accurately diagnosing medical conditions. Additionally, by avoiding artificial data manipulation, we aimed to enhance the model's generalizability and robustness and improve the interpretability and trustworthiness of the model's decisions, particularly in the sensitive context of medical applications. The following preprocessing techniques were applied:

- **Resize**: To standardize the input size, we resized every image to 224 × 224 pixels. This resizing was accomplished with the PIL (Pillow) [29]. Adopting the 224 × 224 pixel resolution aligns with the requirements of well-known pretrained models, including VGG, ResNet, and MobileNet. Standardizing the image dimensions is vital for maintaining consistent operation across all images through uniform convolutional and pooling steps. This uniformity is crucial for generating feature maps of equal dimensions, a necessity for efficient batch processing in model training.
- **Normalization**:Normalization is an important preprocessing step in machine learning that adjusts the RGB pixel values of images to a uniform range, enhancing the efficiency of learning algorithms. To bring all RGB pixel values to a standardized range, normalization was applied. The general equation for min-max normalization is [30]:

$$p' = \frac{p - \min(p)}{\max(p) - \min(p)} \tag{1}$$

TABLE 1. Comparative distribution of training and testing instances across MSID and EMSID datasets.

Classes	MSID Dataset		EMSID Dataset	
	Training	Testing	Training	Testing
Chickenpox	127	32	127	32
Eczema	-	-	83	21
Lupus	-	-	124	31
Measles	73	19	73	19
Molluscum Contagiosum	-	-	78	20
Monkeypox	223	56	223	56
Normal	234	59	234	59
Scabies	-	-	84	21

For 8-bit RGB images, where each color channel has pixel values ranging from 0 to 255, this equation simplifies to:

$$p' = \frac{p}{255} \tag{2}$$

Here, p represents the original pixel value in each color channel, and p' is the normalized value. Normalization serves multiple purposes. Firstly, it scales all features, in this case, RGB pixel values, to a common range, facilitating more efficient learning. Secondly, many activation functions like the sigmoid or hyperbolic tangent (tanh) are designed to operate over a limited range of input values, such as $0 - 1$ or -1 to 1. Normalization helps keep the activations and gradients within a manageable range, aiding in faster and more stable convergence during the training phase.

- **Class Imbalance**: Given the imbalanced nature of the dataset, we carefully stratified the train-test splits to ensure that each subset had a proportional representation of each class.

These preprocessing steps were crucial in preparing the dataset for effective training and evaluation of deep learning models.

C. DATASET SPLITTING

We mainly evaluated our method using the MSID and EMSID datasets. Both dataset were divided using the holdout method [31]. Specifically, 80% of the instances from each class were allocated to the training set, while the remaining 20% were used for the testing set.

The detailed distribution of instances of both dataset across the training and testing sets are outlined in Table 1. This division ensures that the model has a sufficient number of examples for training while also providing an independent set for performance evaluation.

D. APPLIED AND PROPOSED MODELS

In addressing the task of skin disease classification, we propose an attention-empowered MobileNetV2 architecture as shown in fig 4. This architecture is particularly suited for deployment on edge devices due to the lightweight nature of MobileNetV2. Additionally, we have applied various state-of-the-art architectures for comparison purposes. Table 2

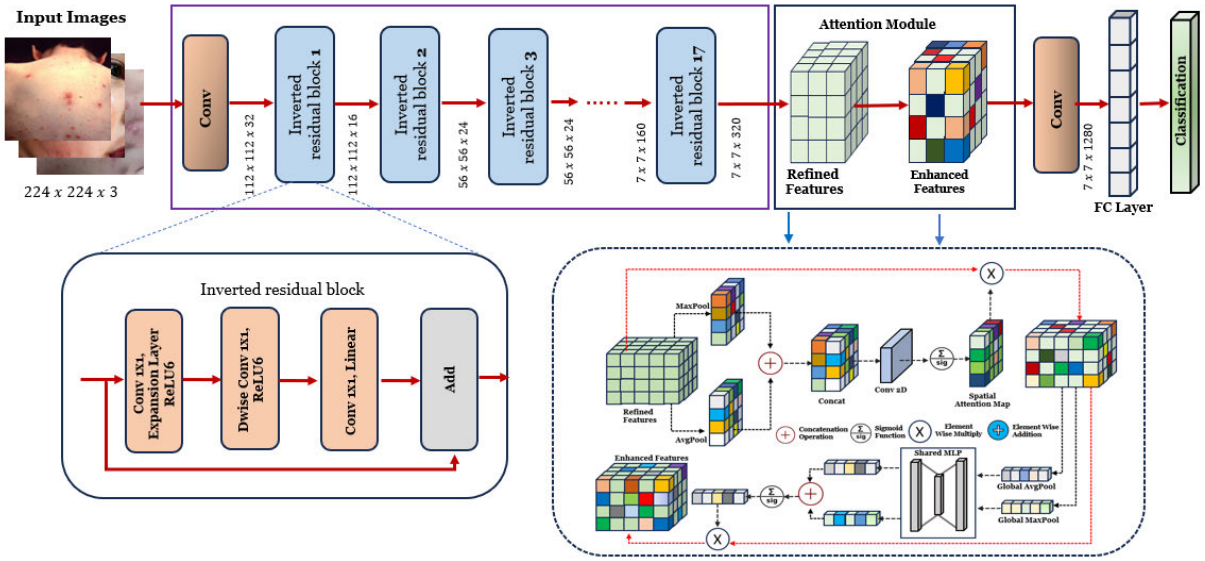

FIGURE 4. Proposed attention empowered MobileNetv2 architecture for skin diseases classification.

TABLE 2. Comparison of evaluated models based on total number of default parameters.

Model	Total Parameters (M)
VGG-19 [32]	143.7
ResNet-152 [33]	60.2
MobileNetV2 [34]	3.5
AlexNet [35]	61.1
GoogLeNet [36]	6.6
ShuffleNetV2 [37]	1.4
Attention Empowered MobileNetV2	3.7

displays the methods applied and compares their total number of parameters. These models are recognized for their effectiveness in image classification tasks. For each deep learning model shown in table 2, we adopted a transfer learning [8] approach. The transfer learning strategy allows us to capitalize on the rich feature representations these models have learned from extensive datasets. This is particularly beneficial in addressing challenges such as data scarcity and computational limitations. It also facilitates rapid prototyping and ensures stable convergence during training. Next, we discuss the detailed architecture of attention-empowered MobileNetV2.

E. ATTENTION EMPOWERED MobileNetV2: A DETAILED ARCHITECTURE

The MobileNetV2 empowered With attention architecture is ingeniously designed to classify skin diseases into eight distinct classes. This architecture leverages a combination of a pre-trained MobileNetV2 model for initial feature extraction. After that, an attention module is applied to focus on the most vital features for distinguishing between various skin conditions. The attention module is comprised of a spatial attention block and a channel attention block. The spatial attention module [38] emphasizes the most significant regions within the image. Then, the channel attention Module [39] is integrated to refine the model's focus further. This attention module concentrates on the various feature channels, enhancing the model's sensitivity to important textural and pattern-based information that might be critical for accurate disease classification. The synergy of these attention mechanisms ensures that the most relevant features are highlighted for the subsequent classification stages. Finally, the architecture culminates in a Fully Connected (FC) layer, synthesizing the extracted and emphasized features to make the final classification decision. In the following section, we provide a detailed description of each architecture component, elucidating their roles and contributions to the overall effectiveness of the model.

1) FEATURE EXTRACTION

MobileNetV2 serves as the feature extraction component in the architecture, optimized particularly for efficiency in mobile and embedded vision applications and edge devices. This efficiency is paramount for real-time skin disease classification tasks where rapid diagnosis is essential, especially for contagious conditions like monkeypox, chickenpox, and measles. Given that some of these diseases are contagious, there is a critical need for fast, accurate diagnosis to initiate timely treatment and isolation procedures [40]. Deploying this model on mobile or edge devices can serve as a frontline diagnostic tool, enabling quick decision-making in various settings, including clinics, schools, and remote locations where computational resources might be limited. The architecture of MobileNetV2 includes two essential techniques: depthwise separable convolutions and inverted residuals. These techniques contribute to the model's efficiency, making it well-suited for real-time analysis on resource-constrained

devices. Next, we describe the two pivotal components of the MobileNetV2.

a: DEPTHWISE SEPARABLE CONVOLUTIONS

MobileNetV2 enhances efficiency by using depthwise separable convolutions, which separate the convolutional process into two layers, reducing computations and parameters. The first layer, depthwise convolution, independently applies a filter to each input channel. The second layer, pointwise convolution, combines these individual channel outputs into new feature maps using 1×1 convolutional filters. This process allows the network to mix information from the different channels, enabling learning more complex features. This process can be represented as follows:

1) **Depthwise Convolution**: Applies a filter f_d independently to each channel of the input X. For each channel c, the output D is computed as [41]:

$$(D_c)_{i,j} = \sum_m \sum_n (f_{d,c})_{m,n} \cdot X_{c,i+m,j+n}, \ \forall c \in \{1, \ldots, C\} \tag{3}$$

Here, $(D_c)_{i,j}$ represents the output for channel c at position (i, j). The filter f_d is applied independently to each channel c of the input.

2) **Pointwise Convolution**: Combines the depthwise outputs using a 1×1 convolution f_p, resulting in the final output Y as [41]:

$$Y_{i,j,k} = \sum_{m=1}^{C} (f_p)_{k,m} \cdot D_{i,j,m}, \ \forall k \in \{1, \ldots, C'\} \tag{4}$$

Here, $Y_{i,j,k}$ is the final output, combining the depthwise outputs. C is the number of channels in the input, and C' is the number of output channels from the pointwise convolution.

The depthwise separable convolution effectively minimizes computational load while retaining the ability to process complex features. The overall operation can be expressed as:

$$Y_{i,j,k} = \sum_{m=1}^{C} (f_p)_{k,m} \cdot \left(\sum_a \sum_b (f_d)_{a,b} \cdot X_{m,i+a,j+b} \right) \tag{5}$$

Fig. 5 represents the whole process of depthwise separable convolutions.

b: INVERTED RESIDUALS IN MobileNetV2

MobileNetV2 introduces inverted residuals, contrasting with traditional residual networks' bottleneck design of contraction, transformation, and expansion shown in Fig. 6. This approach follows a sequence of expansion, depthwise convolution, and projection, enhancing the model's efficiency and feature preservation. This process can be expressed as:

$$\text{Expansion} \rightarrow \text{Depthwise Convolution} \rightarrow \text{Projection} \tag{6}$$

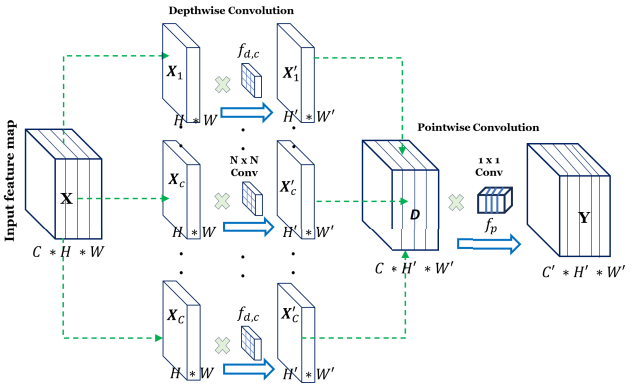

FIGURE 5. The overall architecture of depthwise separable convolutions module.

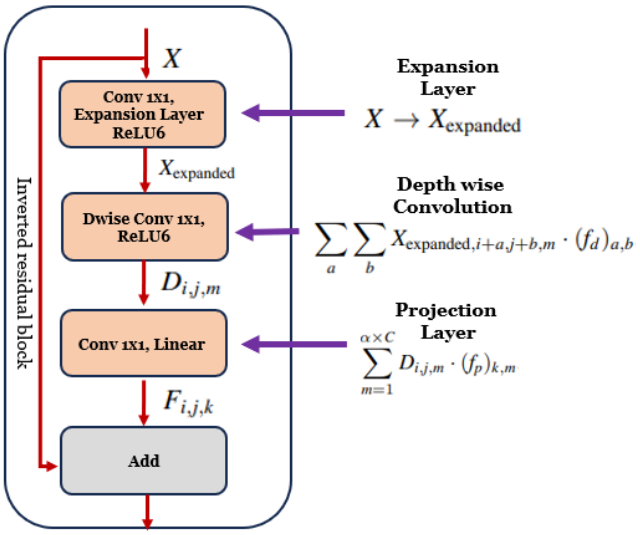

FIGURE 6. The overall architecture of inverted residual module.

This sequence in MobileNetV2 inverts the conventional residual block approach, focusing on first expanding and then refining the feature channels for enhanced efficiency and feature representation. The details of these processes are described as follows:

1) **Expansion** The first layer in an inverted residual block is responsible for expanding the number of channels in the feature map. If C is the number of channels in the input feature map X, then after the expansion layer, the number of channels becomes $\alpha \times C$, where α is the expansion factor (usually greater than 1). This operation can be represented as:

$$X \rightarrow X_{\text{expanded}} \quad (\text{Shape: } \alpha \times C \times H \times W) \tag{7}$$

2) **Depthwise Convolution** A depthwise convolution operation is performed on the expanded feature map. This results in a new feature map D with the same number of channels $\alpha \times C$ but typically with reduced spatial dimensions $H' \times W'$. The operation can be

mathematically represented as [41]:

$$D_{i,j,m} = \sum_{a} \sum_{b} (f_d)_{a,b} \cdot X_{\text{expanded},i+a,j+b,m}, \quad \forall m \in \{1, \ldots, \alpha \times C\} \quad (8)$$

3) **Projection** Finally, a 1×1 convolution (pointwise convolution) is applied to project the feature map back to a lower-dimensional space with C' channels. This operation can be represented as [41]:

$$F_{i,j,k} = \sum_{m=1}^{\alpha \times C} D_{i,j,m} \cdot (f_p)_{k,m}, \quad \forall k \in \{1, \ldots, C'\} \quad (9)$$

where F is the final feature map with dimensions $C' \times H' \times W'$.

2) SPATIAL ATTENTION MODULE

The spatial attention module plays a vital role, focusing specifically on enhancing the influence of significant spatial regions within the feature map, which is particularly beneficial for skin disease classification [42]. It generates an attention map, effectively modulating the original feature map F to highlight crucial areas for accurate disease identification. In the context of skin disease classification, certain regions of a skin image contain critical information about the disease, such as the texture, color, or shape of a lesion. The spatial attention module helps the network to pay more attention to these informative regions, improving the model's ability to differentiate between various skin conditions. This is especially important in cases where the disease manifests with subtle visual cues, making it challenging to diagnose without a focused examination of the affected area. The spatial attention module operates through several steps to achieve this enhancement of relevant spatial features:

1) **Feature Aggregation**: The module first aggregates information across the feature map F to generate a descriptor that captures the global context of the input. This is done using pooling operations. The aim is to create a compact representation of the feature map that highlights the most salient features for the task. The spatial attention uses the following two pooling operations.
   a) **Average Pooling**: The average pooling aims to capture the central tendency of the activations across the channel dimension for each spatial location (h, w). Mathematically, for each spatial location (h, w), the average is computed as follows:

$$\xi_{h,w}^{avg} = \frac{1}{C'} \sum_{c=1}^{C'} F_{c,h,w} \quad (10)$$

   This operation condenses the multi-channel feature map F into a single-channel feature map $\xi_{1,H,W}^{Avg}$, effectively emphasizing areas where the average activation is high. Note that, the notation $\xi_{h,w}^{avg}$ and $\xi_{1,H,W}^{Avg}$ are different. $\xi_{h,w}^{avg}$ represents a single value whereas $\xi_{1,H,W}^{Avg}$ represents the whole whole feature-map.
   b) **Max Pooling**: The max pool operation, on the other hand, captures the peak activations across the channel dimension for each spatial location (h, w). This is expressed mathematically as:

$$\alpha_{h,w}^{max} = \max_{c=1}^{C'} F_{c,h,w} \quad (11)$$

   Similar to the average computation, this operation produces a single-channel feature map $\alpha_{1,H,W}^{Max}$ that emphasizes regions of the original feature map where at least one channel has high activation.

The rationale for using both average and maximum computations is to capture both the general and extreme characteristics of the feature map, thereby providing a more robust basis for the attention mechanism. After computing the average $\xi_{1,H,W}^{Avg}$ and maximum $\alpha_{1,H,W}^{Max}$ feature maps, they are concatenated along the channel dimension to form a new tensor Υ^{concat}. This tensor has a shape of $2 \times H \times W$, combining both average and maximum statistics for further processing. The concatenation is defined as:

$$\Upsilon_{2,H,W}^{\text{concat}} = [\xi_{1,H,W}^{Avg}, \alpha_{1,H,W}^{Max}] \quad (12)$$

This succinctly captures the process of creating a comprehensive feature representation by combining average and max pooled maps.

2) **Attention Map Generation**: The attention map is generated using the concatenated feature map $\Upsilon_{2,H,W}^{\text{concat}}$ through a convolution operation, resulting in Ψ. This convolution integrates both average and max activations for spatial emphasis, using a $k \times k$ kernel to produce an output of dimensions $1 \times H \times W$. The convolution operation is defined as:

$$\Psi_{h,w} = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} (f)_{m,n} \cdot \Upsilon_{h+m,w+n}^{\text{concat}}, \ \forall h, w \quad (13)$$

A sigmoid activation function σ normalizes the convolution output, resulting in the attention map $\Phi_{h,w}$ with values between 0 and 1:

$$\Phi_{h,w} = \sigma(\Psi_{h,w}) = \frac{1}{1 + \exp(-\Psi_{h,w})}, \ \forall h, w \quad (14)$$

3) **Feature Modulation**: The original feature map F is modulated by the attention map, $\Phi_{h,w}$ through element-wise multiplication, enhancing important regions and suppressing others. This can be defined as [43]:

$$F_{c,h,w}^{\text{new}} = F_{c,h,w} \times \Phi_{h,w}, \ \forall c, h, w \quad (15)$$

After the spatial attention, we will get the feature map $F_{C,H,W}^{\text{new}}$ from the input feature map F. This feature map $F_{C,H,W}^{\text{new}}$ will be further focused by using the channel attention module.

3) CHANNEL ATTENTION MODULE

The channel attention module is a critical component specifically aimed at enhancing the model's sensitivity to pivotal channel-wise features in the feature map $F^{\text{new}}_{C,H,W}$. This is particularly relevant in skin disease classification, where distinct channel characteristics may encode vital diagnostic information such as color variations and texture details. By amplifying the significant channels in the feature map, the channel attention module aids the network in focusing on these critical aspects, thus improving its ability to distinguish between different skin conditions.

The channel attention module operates through a series of steps to augment the relevant channel features:

1) **Channel-Wise Feature Aggregation**: Initially, the module computes channel-wise statistics across the spatial dimensions of the feature map $F^{\text{new}}_{C,H,W}$. This process involves pooling operations to condense the spatial information into a channel-wise descriptor. The goal here is to distill the essential spatial features of each channel into a compact form. The module employs two pooling strategies:
   a) **Global Average Pooling**: This operation computes the average of spatial features for each channel c, capturing the essence of each channel's contribution. Mathematically, the global average for channel c is defined as [44]:

$$\gamma_c^{avg} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} F^{\text{new}}_{c,h,w} \tag{16}$$

   This results in a vector $\Gamma^{\text{Avg}}_{C,1,1}$, emphasizing the channels based on their average spatial activation.
   b) **Max Pooling**: This pooling captures the maximum activation across the spatial dimensions for each channel, focusing on the most prominent features. The mathematical expression is [44]:

$$\beta_c^{max} = \max_{h=1}^{H} \max_{w=1}^{W} F^{\text{new}}_{c,h,w} \tag{17}$$

   It produces a vector $\Omega^{\text{Max}}_{C,1,1}$ that highlights channels with peak spatial activation.

   The combination of both average and max pooling provides a comprehensive view of each channel's spatial features, serving as the basis for the attention mechanism.
2) **Shared MLP Network**: The spatial context descriptors are processed through a shared Multi-Layer Perceptron (MLP) network. This network has one hidden layer with an activation size of $R^{C/r \times 1 \times 1}$, where r is the reduction ratio. The MLP weights are $W_0 \in \mathbb{R}^{C/r \times C}$ and $W_1 \in \mathbb{R}^{C \times C/r}$, with ReLU activation following W_0.
3) **Channel Attention Map Generation**: The channel attention map $M_c \in \mathbb{R}^{C \times 1 \times 1}$ is formulated by element-wise summation of the MLP outputs for both descriptors [44]:

$$M_{C,1,1} = \sigma(W_1(W_0(\Gamma^{\text{Avg}}_{C,1,1})) + W_1(W_0(\Omega^{\text{Max}}_{C,1,1}))), \tag{18}$$

   where σ is the sigmoid function. This map dynamically scales the channels in $F^{\text{new}}_{C,H,W}$, thereby emphasizing those most relevant for the task of skin disease classification. Then the $M_{C,1,1}$ could be element wise multiplied to $F^{\text{new}}_{C,H,W}$ to generate the enhanced features denoted by $F^{\text{Enhanced}}_{C,H,W}$. The details of the whole attention module are given in fig. 7.

Integrating the Channel Attention Module into our architecture significantly improves the model's ability to discern crucial channel-wise features, thereby boosting its performance in diagnosing various skin conditions.

4) GLOBAL AVERAGE POOLING AND CLASSIFICATION

The enhanced attention map $F^{\text{Enhanced}}_{c,h,w}$ is first processed by a convolution layer, yielding $\chi_{c,h,w}$. This is followed by global average pooling:

$$Z_c = \frac{1}{H' \times W'} \sum_{h=1}^{H'} \sum_{w=1}^{W'} \chi_{c,h,w} \; \forall c \in \{1, \ldots, C'\} \tag{19}$$

The pooled vector Z is then fed into a fully connected layer for classification:

$$\Lambda = W \cdot Z + b \tag{20}$$

Table 4 shows the details of each layer, including the input-output dimensions, in the proposed model.

5) LOSS FUNCTION: CROSS-ENTROPY LOSS

We have employed cross-entropy loss as the loss function of the classification model, particularly because it is well-suited for multi-class classification problems such as skin disease diagnosis [1], [14]. The cross-entropy loss L for a single sample is defined as [21]:

$$L = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{21}$$

where y is the true label vector, $\hat{y}$ is the predicted probability distribution, and C is the number of classes. The loss is calculated for each sample and averaged over the entire batch to update the model parameters.

6) OPTIMIZER: ADAM OPTIMIZER

The Adam optimizer is used to adaptively adjust the learning rates during training. This optimizer is particularly useful for medical diagnostic tasks where a balance between fast convergence and model stability is critical. The weight update rule in Adam can be mathematically represented as [45]:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \times \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{22}$$

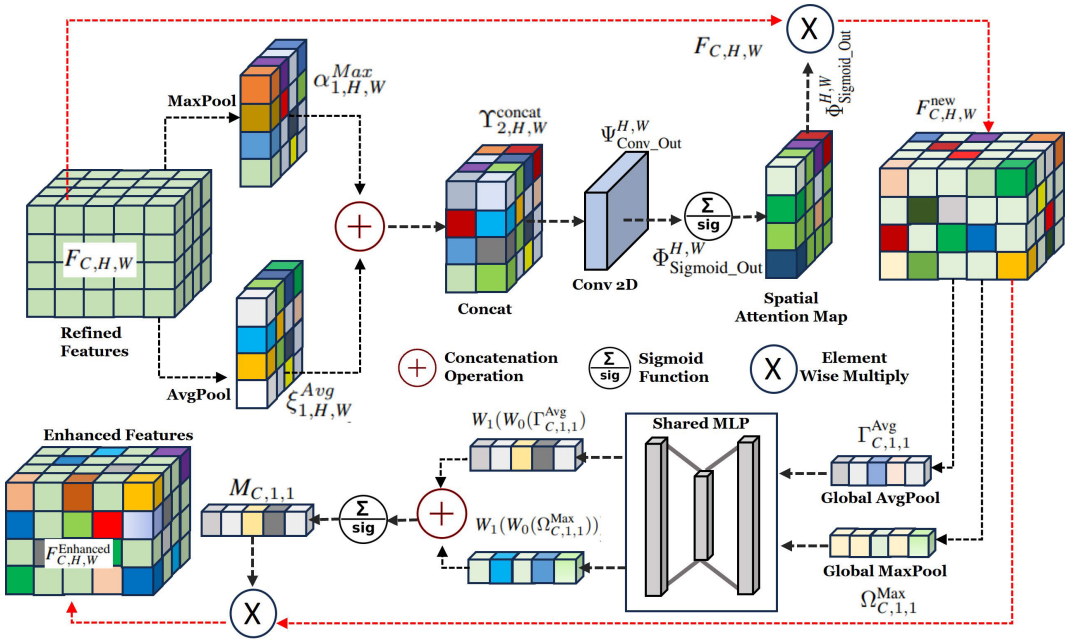

FIGURE 7. Detail architecture of attention module.

Here, θ_{old} and θ_{new} are the old and new weights, respectively. η is the learning rate, m_t and v_t are the first and second-moment estimates, and ϵ is a small constant to prevent division by zero. m_t and v_t can be represented by the following equations [45]:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla_\theta L \tag{23}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla_\theta L)^2 \tag{24}$$

where β_1 and β_2 are the exponential decay rates for the first and second moment estimates, and $\nabla_\theta L$ is the gradient of the loss function with respect to the weights. This optimizer is well-suited for problems that are large in terms of data and/or parameters, making it an excellent choice for complex architectures used in medical diagnostics.

IV. OPERATIONAL WORKFLOW

This section delineates a systematic description of the operational workflow for the early detection of monkeypox. A comprehensive overview of the entire framework is depicted in Figure 8.

1) **Step 1: Model Acquisition**
   Initially, patients are required to download the pre-trained model from a designated server. The key parameters governing the training of the proposed model are detailed in Table 3.
2) **Step 2: Image Capture**
   Subsequently, patients have to capture images of the affected areas. This step is crucial for ensuring that the input data to the model is of high quality and relevance.
3) **Step 3: Image Processing and Classification**
   The captured images are then processed and input into the trained model for classification. This process involves the model making a diagnostic decision based on the visual evidence presented in the images.
4) **Step 4: Interpretability and Verification**
   Utilizing techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME), patients can verify whether the model's focus aligns with the clinically relevant areas. If the focus is misaligned, patients are advised to recapture the image.
5) **Step 5: Decision Support**
   Upon confirmation that the model is focusing on the correct areas, patients can then proceed to make informed decisions based on the model's diagnosis.

TABLE 3. Summary of the key parameters of attention-based MobileNetV2.

Parameter	Value/Description
Model Architecture	Attention-based MobileNetV2
Programming Language	Python
Deep Learning Framework	PyTorch 2.0.1
Input Image Size	224x224
Preprocessing Requirements	Image resizing, normalization
Training Epochs	100
Learning Rate	0.0001
Optimizer	Adam
Loss Function	Categorical Crossentropy

V. BASELINES AND EVALUATION METRICS

A. BASELINE MODELS

We have used the following baselines for comparing the performance and efficiency of various deep learning models in the context of skin disease classification, ensuring a comprehensive evaluation against state-of-the-art architectures. The baselines are as follows:

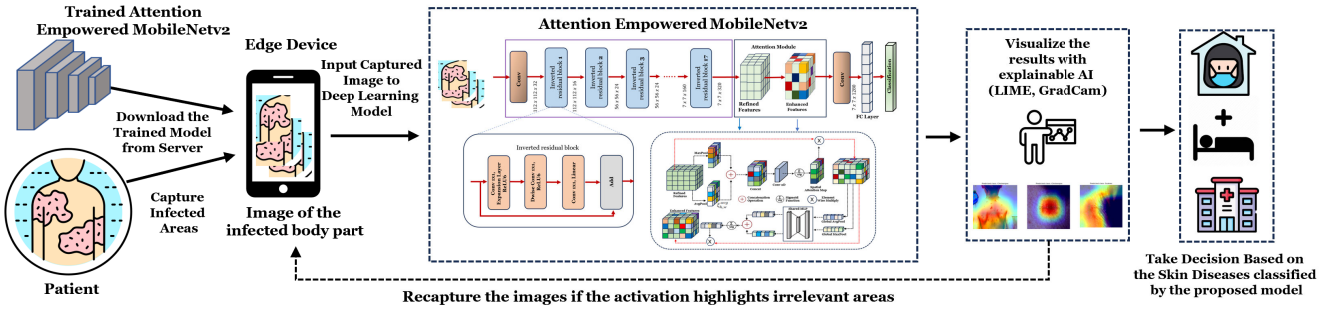

FIGURE 8. Detailed operational workflow for the early detection of monkeypox.

TABLE 4. Layer-wise architectural overview of the enhanced MobileNetV2 with attention modules.

Layer (type)	Input Shape	Output Shape
MobileNetV2 With Attention Module	[32, 3, 224, 224]	[32, 8]
Sequential		
Conv2dNormActivation (0)	[32, 3, 224, 224]	[32, 32, 112, 112]
BatchNorm2d (1)	[32, 32, 112, 112]	[32, 32, 112, 112]
ReLU6 (2)	[32, 32, 112, 112]	[32, 32, 112, 112]
InvertedResidual Blocks (1-17)		
Inverted Residual (1)	[32, 32, 112, 112]	[32, 16, 112, 112]
Inverted Residual (2)	[32, 16, 112, 112]	[32, 24, 56, 56]
Inverted Residual (3)	[32, 24, 56, 56]	[32, 24, 56, 56]
Inverted Residual (4)	[32, 24, 56, 56]	[32, 32, 28, 28]
Inverted Residual (5)	[32, 32, 28, 28]	[32, 32, 28, 28]
Inverted Residual (6)	[32, 32, 28, 28]	[32, 64, 14, 14]
Inverted Residual (7)	[32, 64, 14, 14]	[32, 64, 14, 14]
Inverted Residual (8)	[32, 64, 14, 14]	[32, 96, 14, 14]
Inverted Residual (9)	[32, 96, 14, 14]	[32, 160, 7, 7]
Inverted Residual (10)	[32, 160, 7, 7]	[32, 160, 7, 7]
Inverted Residual (11)	[32, 160, 7, 7]	[32, 160, 7, 7]
Inverted Residual (12)	[32, 160, 7, 7]	[32, 320, 7, 7]
Inverted Residual (13)	[32, 320, 7, 7]	[32, 1280, 7, 7]
Inverted Residual (14)	[32, 160, 7, 7]	[32, 160, 7, 7]
Inverted Residual (15)	[32, 160, 7, 7]	[32, 160, 7, 7]
Inverted Residual (16)	[32, 160, 7, 7]	[32, 320, 7, 7]
Inverted Residual (17)	[32, 320, 7, 7]	[32, 1280, 7, 7]
Attention Module (spatial_attention)		
Conv2d (conv1)	[32, 2, 7, 7]	[32, 1, 7, 7]
Sigmoid (sigmoid)	[32, 1, 7, 7]	[32, 1, 7, 7]
Attention Module (Channel_attention)		
Adaptive Avg Pool 2d	[32, 320, 7, 7]	[32, 320, 1, 1]
Sequential (fc)	[32, 320]	[32, 320]
Adaptive Max Pool 2d	[32, 320, 7, 7]	[32, 320, 1, 1]
Sequential (fc)	[32, 320]	[32, 320]
Sequential		
Conv2d (0))	[32, 320, 7, 7]	[32, 1280, 7, 7]
BatchNorm2d (1)	[32, 1280, 7, 7]	[32, 1280, 7, 7]
ReLU6 (2)	[32, 1280, 7, 7]	[32, 1280, 7, 7]
Linear (fc)		
Linear	[32, 1280]	[32, 8]

1) **VGG-19:** VGG-19, a deep convolutional neural network, performs excellently across various image classification tasks. With its 19-layer architecture, VGG-19 can capture intricate patterns in image data, making it a robust choice for our task. We employed the VGG-19 model pre-trained on the ImageNet dataset. Then, it was subsequently fine-tuned with initial weights specifically for classifying skin diseases, including monkeypox, using the EMSID dataset.
2) **ResNet-152:** ResNet-152 introduces the concept of residual learning, which facilitates the training of deep networks by addressing the vanishing gradient problem. This 152-layer network ensures a deep understanding of the image data, which is vital for distinguishing between the various skin conditions present in our dataset. Our study used ResNet-152 pre-trained on ImageNet.
3) **MobileNetV2:** MobileNetV2 is renowned for its efficiency and lightweight nature. It utilizes depthwise separable convolutions, significantly reducing the number of parameters without compromising performance as we have discussed earlier. Given the computational

constraints we aim to adhere to, this feature makes MobileNetV2 particularly suitable for our study.

4) **AlexNet:** AlexNet, a deep convolutional neural network, marked a significant advancement in deep learning for image recognition tasks. With its 8-layer architecture, AlexNet is known for its breakthrough performance in large-scale image classification challenges. In our study, we incorporated AlexNet pre-trained on the ImageNet dataset. This established a robust base for extracting features as we fine-tuned the network to specifically cater to the intricacies of classifying skin diseases from the EMSID dataset.
5) **GoogleLeNet (Inception v1):** GoogleLeNet, or Inception v1, introduces the inception module, an innovative approach that concatenates feature maps produced by varying-size filters. This architecture allows the network to capture spatial hierarchies in images at different scales, making it highly effective for complex image classification tasks. In our research, we utilized GoogleLeNet pre-trained on the ImageNet dataset. This provided a strong foundation for feature extraction while we fine-tuned the network to align with the nuances of skin disease classification.
6) **ShuffleNetV2:** ShuffleNetV2, an advancement in efficient neural network design, is renowned for its unique approach that optimizes computational efficiency through the use of channel shuffling and pointwise group convolutions. This architecture is specifically engineered to maintain high accuracy while significantly reducing computational complexity and memory usage, making it ideal for deployment on devices with limited resources such as mobile phones or edge devices.

Table 2 presents the weights of the various models used in this study, offering insights into their complexity and resource requirements.

B. EVALUATION METRICS

Traditional metrics such as accuracy, precision, recall, specificity, and F1-Score are commonly used to evaluate classification models. These metrics provide a foundational understanding of model performance, particularly regarding error type (e.g., false positives vs. false negatives). However, they only sometimes offer a complete picture, especially in class imbalances or when the cost of different types of errors varies. To address these limitations and gain a more nuanced understanding of model performance, we supplement these traditional metrics with the following measures:

1) **Cohen's Kappa [46]:** In skin disease classification, accurate and reliable diagnosis is crucial due to the conditions' potential severity and treatment costs. Traditional evaluation metrics such as accuracy, precision, and recall are commonly used but may not suffice due to class imbalances and the varying consequences of misclassification errors. Therefore, we also employ Cohen's Kappa score to provide a more nuanced measure of our model's performance. The measure is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{25}$$

where p_o is the proportion of observed agreements, and p_e is the expected proportion of chance agreements. The kappa statistic is indicative of a classifier's ability to discern between conditions accurately, which is especially important given the similarity in presentation among various skin diseases. A higher value of κ suggests that the classifier's performance is due to its predictive ability rather than chance.

2) **Matthews Correlation Coefficient [47]:** The Matthews Correlation Coefficient (MCC) is a reliable statistical rate that produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP)), proportionally both to the size of positive elements and the size of harmful elements in the dataset. It is particularly informative in binary classification tasks, even when the classes are very different sizes. The MCC is defined as:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP}) \cdot (\text{TP}+\text{FN}) \cdot (\text{TN}+\text{FP}) \cdot (\text{TN}+\text{FN})}} \tag{26}$$

An MCC of +1 represents a perfect prediction, 0 is no better than a random prediction, and -1 indicates total disagreement between prediction and observation.

3) **Youden's J Index [48]:** Youden's J Index is a single statistic that captures the performance of a diagnostic test, summarizing the ROC curve. Unlike accuracy, which does not differentiate between the types of errors, Youden's J incorporates both sensitivity (true positive rate) and specificity (true negative rate) to provide a more robust measure of class imbalance. The index is defined as:

$$\text{J} = \text{sensitivity} + \text{specificity} - 1 \tag{27}$$

where sensitivity is the probability that a test result will be positive when the disease is present (true positive rate), and specificity is the probability that a test result will be negative when the disease is not present (true negative rate). A higher Youden's J Index indicates a better performance of the test to identify positive cases without misclassifying negative cases.

VI. RESULTS ANALYSIS, DISCUSSION AND EXPLAINABILITY

A. PERFORMANCE ON MSID DATASET

We also evaluate the proposed attention-based model on the MSID dataset. Fig. 9 shows the confusion matrix of the

TABLE 5. Evaluating classification efficacy: A detailed metrics analysis of various deep learning models on EMSID.

Model	Precision	Recall	F1-score	Specificity	Sensitivity	Accuracy
VGG-19	0.8372	0.8263	0.8300	0.9794	0.8263	0.8571
ResNet-152	0.8959	**0.9040**	0.8983	0.9880	**0.9040**	0.9147
MobileNetV2	0.8819	0.8860	0.8823	0.9869	0.8860	0.9073
GoogLenet	0.8778	0.8772	0.8755	0.9865	0.8772	0.9035
AlexNet	0.8377	0.8165	0.8240	0.9804	0.8165	0.8649
ShuffleNetV2	0.8080	0.8042	0.8062	0.9780	0.8042	0.8455
Proposed	**0.9048**	0.8942	**0.8984**	**0.9890**	0.8942	**0.9228**

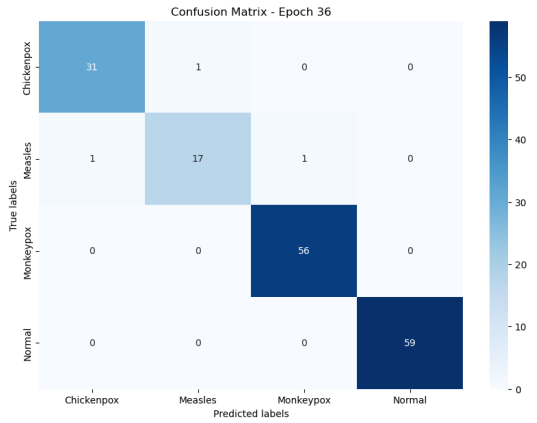

FIGURE 9. Confusion matrix of proposed model on MSID dataset.

proposed model on the MSID dataset. From the figure, we can see that with an overall accuracy of 98.19%, the model shows exceptional capability in correctly classifying the images. It maintains high precision (97.39%), indicating a solid ability to minimize false positives, which is critical in medical diagnostics. The recall or sensitivity of 96.59% reflects the model's proficiency in correctly identifying positive cases, ensuring that few cases are missed. The F1-score, at 96.97%, indicates a balanced performance between precision and recall. The specificity of 99.42% further underscores the model's strength in accurately identifying negative cases, which is essential for reducing false alarms.

B. RESULT ANALYSIS ON EMSID

After completing the training and evaluation, the results show that the proposed model can accurately classify different skin lesions. The proposed model's performance was satisfactory, showing it can tell the difference between different types of skin lesions. These results confirm that the model is reliable and effective in correctly identifying and classifying skin lesions.

Fig. 10 presents the confusion matrices for different models. The analysis of these matrices reveals that all models exhibit proficiency in detecting monkeypox alongside other skin lesions. However, the proposed attention-based model correctly classified all the 56 monkeypox instances from the test dataset. The class-wise performance of each model is presented in Fig. 11. Table 5 compares precision, recall, F1-score, sensitivity, specificity, and accuracy among the baseline models given in table 2.

We can see from table 5, the proposed model achieved a remarkable precision (0.9048), outshining other baselines. This high precision indicates the model's robustness in accurately identifying positive instances among various skin conditions. In terms of specificity (0.9890) and accuracy (0.9228), the proposed model again led the field, demonstrating its ability to correctly classify negative instances and maintain overall accuracy across diverse conditions.

In contrast, while ResNet-152 showed commendable performance, particularly in recall (0.9040), it slightly lagged behind from the proposed model in other critical metrics. This highlights the nuanced differences in model performance, where some excel in recognizing true positive cases while others balance precision and recall more effectively.

The standard MobileNetV2, without the attention mechanisms, also demonstrated solid performance but was clearly outperformed by the proposed model in terms of precision and accuracy. This underlines the significant impact of incorporating attention mechanisms into the network architecture. VGG-19 and AlexNet, while moderately effective, were outclassed by the more advanced models, particularly in terms of specificity and accuracy. The ShuffleNetV2, which has the lowest number of parameters, also showed the worst performance across all evaluated metrics. This underscores the trade-off between model complexity and efficiency, with ShuffleNetV2's streamlined design not capturing as much detailed information as its more complex counterparts, leading to reduced effectiveness in this specific task of skin disease classification.

This comparative analysis highlights the importance of attention mechanisms in enhancing the diagnostic capabilities of neural networks, especially in skin disease images. Integrating spatial and channel attention modules in the MobileNetV2 model has facilitated a more focused and nuanced analysis of dermatological images, leading to more accurate classifications. This advancement is crucial in medical imaging, where the accurate diagnosis of skin diseases can be challenging due to the subtle variations in visual features across different conditions.

Fig. 12 compares the performance of different classifiers for the monkeypox class. The figure shows that the attention-based MobileNetV2 model has demonstrated superior performance for detecting the instances from the monkeypox class from the test dataset. The model has achieved an impressive accuracy of 98.07%, with perfect

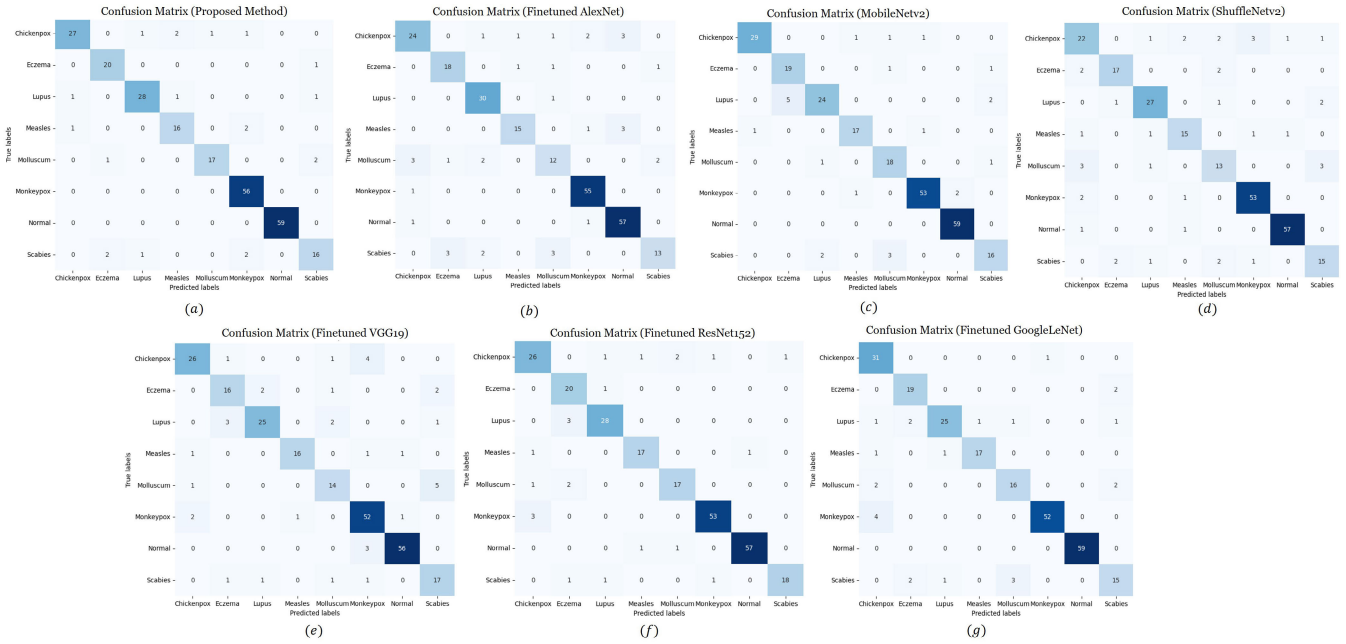

FIGURE 10. **Confusion matrices for skin lesion classification: (a) Proposed model, (b) AlexNet, (c) MobileNetV2, (d) ShuffleNetV2 (e) VGG-19, (f) ResNet-152 and (g) GoogLeNet models, highlighting comparative performance.**

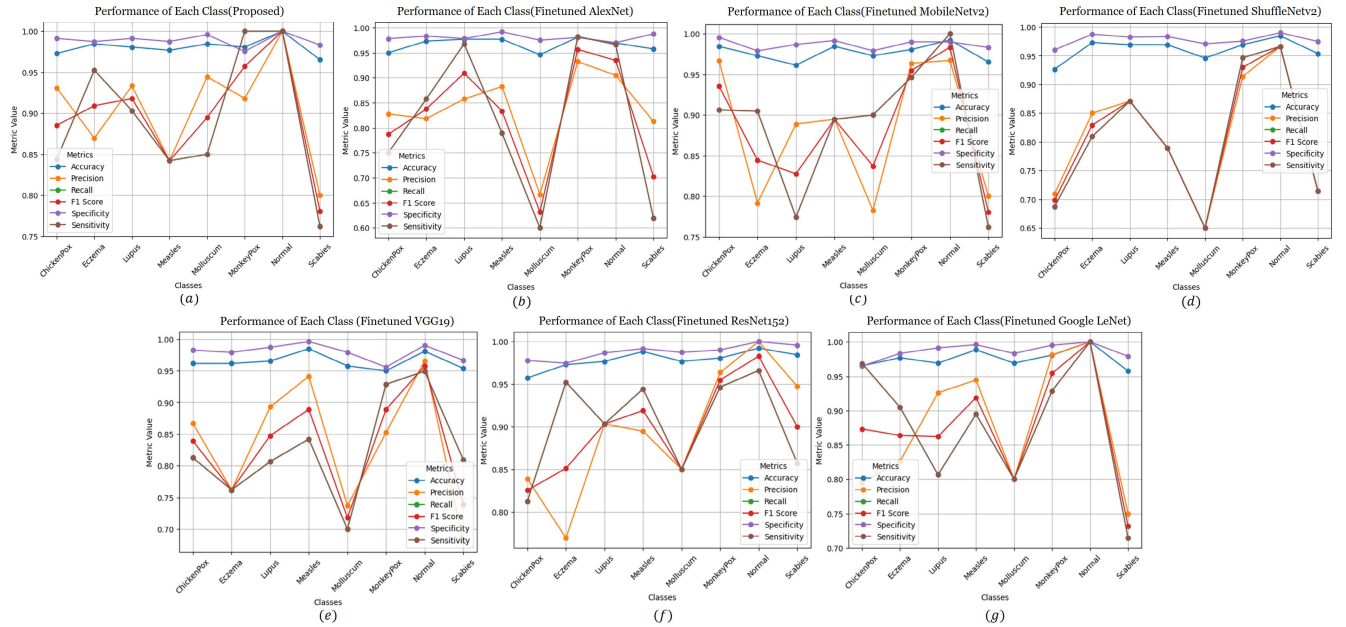

FIGURE 11. **Comprehensive class-wise evaluation of skin lesion detection models: (a) Proposed model, (b) AlexNet, (c) MobileNetV2, (d) ShuffleNetV2 (e) VGG-19, (f) ResNet-152, and (g) GoogLeNet - A comparative analysis highlighting accuracy, precision, recall, F1-score, specificity and sensitivity.**

sensitivity, indicating its ability to detect all true monkeypox cases without any false negatives. The precision of the model stands at 91.8%, which, combined with a recall (sensitivity) of 100%, yields an F1-score of 97.39%. Furthermore, the specificity of 97.53% ensures that the model is highly effective at correctly identifying non-monkeypox cases, thereby minimizing the risk of false positives. In comparison, other models like the MobileNetV2, VGG-19, and ResNet-152 have shown commendable results; however, they fall short of the benchmarks the attention-based MobileNetV2 model sets. Notably, the standard MobileNetV2 and ResNet-152 models have recorded an accuracy of 98.06%, with F1-scores nearing 95.50%. VGG-19, while still performing well, lags slightly behind in recall and F1-score, indicating a greater tendency

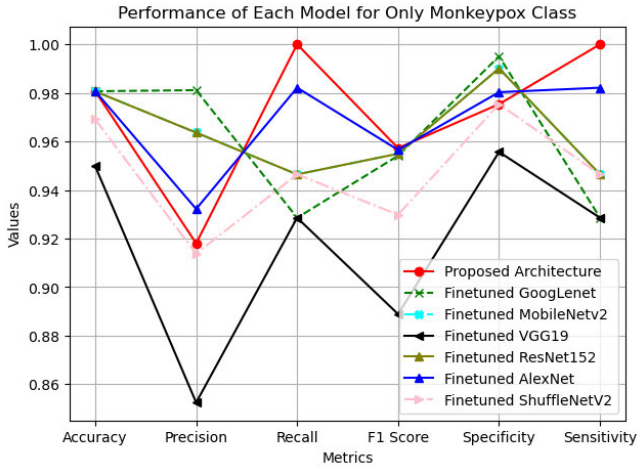

FIGURE 12. Comparative analysis of different models for monkeypox classification, evaluating accuracy, precision, recall, F1-score, specificity, and sensitivity.

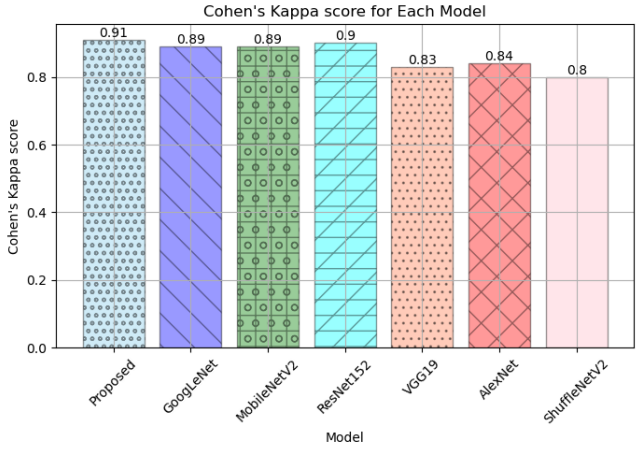

FIGURE 13. Cohen's kappa score comparison across models for enhanced reliability assessment.

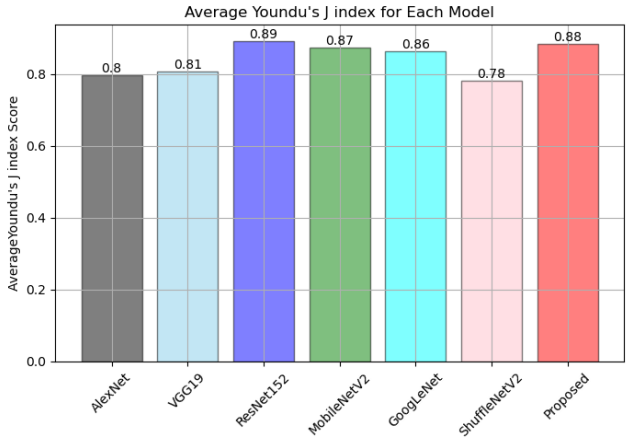

FIGURE 14. Comparison of Youndu's J score among different models.

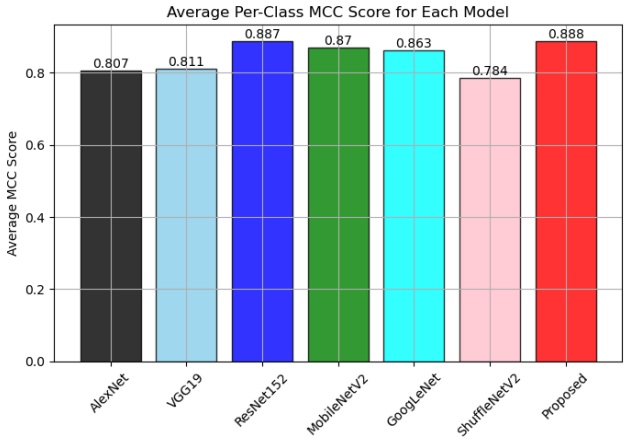

FIGURE 15. Comparison of MCC score among different models.

to miss actual monkeypox cases compared to the proposed model.

Fig. 13 shows Cohen's kappa scores for each evaluated model, illustrating the degree of agreement between the models' predictions and the ground truth for the classification of dermatological diseases. The proposed attention-based MobileNetV2 model demonstrated a remarkable Cohen's kappa score of 0.91, reflecting a substantial agreement with the ground truth. This score underscores the model's precision in discerning the nuances of dermatological imaging. Traditional architectures like ShufflenetV2, GoogLeNet, MobileNetV2, and ResNet-152 yielded kappa scores of 0.80, 0.89, 0.89, and 0.90, respectively. These scores, while indicative of high agreement, still leave room for improvement, which the attention mechanisms seem to address effectively. Notably, VGG-19 and AlexNet, despite their depth and complexity, obtained a kappa score of 0.83 and 0.84, respectively, suggesting that their performance, in terms of agreement with clinical diagnoses, is less reliable compared to the other models in the context of this study.

Fig. 14 shows Youden's J scores for each evaluated model. The figure shows that ResNet-152 leads the pack with a score of 0.892, indicating its superior ability to identify positive and negative cases accurately. However, the proposed model stands out with a high score of 0.883, highlighting its specialized features' effectiveness in accurately diagnosing skin diseases. The Matthews Correlation Coefficient (MCC) scores for all the classifiers have been shown in Fig. 15. The figure illustrates that AlexNet, VGG-19 and ShuffleNetV2, with an MCC of 0.807, 0.811 and 0.784 show reasonable capability yet indicate potential for improvement in balanced classification accuracy. VGG-19 marginally surpasses AlexNet and ShuffleNetV2 with an MCC of 0.811, reflecting a slightly better balance between true and false positives and negatives. ResNet-152 stands out with an MCC of 0.887, showcasing its superior performance in making balanced and accurate predictions. MobileNetV2, with an MCC of 0.870, demonstrates robust performance, balancing efficiency and accuracy, while GoogLeNet scores 0.863, suggesting competent performance albeit slightly

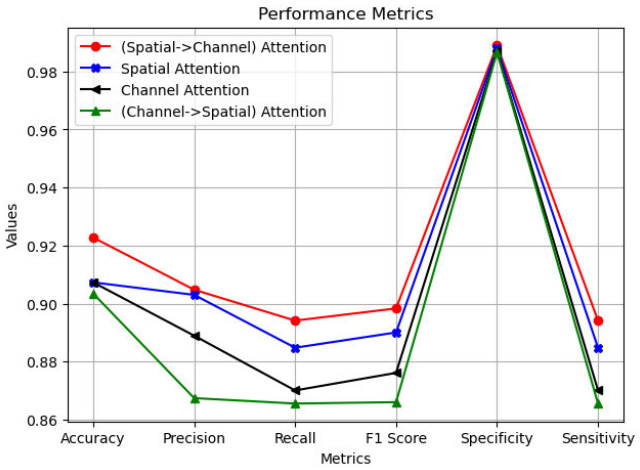

FIGURE 16. Detailed comparative analysis of sequential spatial-to-channel and channel-to-spatial attention mechanisms versus singular attention models on precision, recall, F1-score, specificity, sensitivity, and accuracy in skin diseases classification.

outperformed by more advanced models. The Proposed model, enhanced with specialized features, achieves the highest MCC of 0.888, underscoring the effectiveness of its design in accurately classifying diverse skin conditions. These scores highlight the importance of model architecture and specialized features in achieving a balanced and accurate classification in complex tasks like skin disease diagnosis.

C. ABLATION STUDY

We conduct an ablation study to assess the impact of different combinations of spatial and channel attention. The outcomes of this study, illustrating the effects of varying spatial and channel attention combinations, are presented in figure 16 and table 6. A summary of the findings derived from figure 16 is as follows:

- **Spatial → Channel Attention (Proposed)**: Achieving the highest scores across all metrics, with an accuracy of 0.923, precision of 0.905, recall of 0.894, F1-score of 0.898, specificity of 0.989, and sensitivity of 0.894. This configuration excels in performance, likely due to its sequential focus on spatial and then channel-wise information, leading to effective classification.
- **Only Spatial Attention**: Shows a slight decrease in performance compared to the full model, indicating that while spatial attention is crucial for focusing on relevant image regions, the lack of channel attention results in a reduced ability to distinguish similar diseases.
- **Only Channel Attention**: This configuration sees further drops in performance metrics, suggesting that while channel attention is important for emphasizing features, it is less effective than spatial attention in isolating key image areas.
- **Channel Attention → Spatial Attention**: Exhibits the lowest performance among the configurations, possibly because prioritizing channel features before spatial features lead to less effective initial analysis, which spatial attention struggles to rectify.

These results indicate that the combination and sequence of spatial and channel attention mechanisms significantly impact the model's accuracy in classifying skin diseases. With spatial followed by channel attention, the proposed model provides a balanced focus on relevant image areas and crucial features, leading to superior performance. Table 6 presents a comparative analysis utilizing three crucial metrics for evaluation: the Matthews Correlation Coefficient (MCC) Score, Youden's J Score, and Cohen's Kappa Score.

The proposed model emerges as the most effective configuration, applying spatial attention followed by channel attention. This model excels across all metrics, achieving an average MCC score of 0.91, Youden's J score of 0.88, and Cohen's Kappa score of 0.91. This superior performance accentuates the advantage of first focusing on spatial features to identify relevant areas within the images, followed by an in-depth analysis of channel-wise details, enhancing the model's diagnostic accuracy.

On the other hand, models employing a singular type of attention mechanism, either "Channel Attention" or "Spatial Attention," demonstrate commendable performance but do not reach the heights of the combined approach. With MCC scores of 0.89 and 0.90, respectively, and closely matching Youden's J and Cohen's Kappa scores, these models underscore the significant contributions of each attention mechanism. However, they simultaneously suggest that a synergistic approach is more conducive to achieving heightened accuracy. Intriguingly, reversing the attention application (Channel → Spatial) sequence records the lowest scores across all metrics. With an MCC of 0.88, Youden's J of 0.85, and Cohen's Kappa of 0.89, this configuration's relatively diminished performance indicates that prioritizing channel features before spatial features may be less efficacious for skin disease classification.

We perform another ablation study to see the effects of the attention mechanism (spatial → channel attention) at different layers of the MobileNetV2. Table 7 shows the result comparison. According to this table, applying attention after the 17th layer yields the best performance across all metrics. This is likely because layer 17 is the final inverted residual layer, followed by just a convolutional layer (see fig 4). The concluding convolutional layer in architectures like MobileNetV2 plays a pivotal role in the final feature refinement prior to classification. By introducing attention right before this layer, the network is enabled to concentrate on and enhance the most vital features for distinguishing between various skin conditions. This strategy proves particularly effective as it precedes the ultimate decision-making stage in the process.

D. DISCUSSION ON COMPARATIVE EFFECTIVENESS AND EFFICIENCY WITH EXISTING LITERATURE

Table 8 highlights the competitive accuracy of our attention-based MobileNetV2 model in detecting monkeypox from skin lesion images. It is noteworthy that certain methods

TABLE 6. Detailed comparative analysis of sequential spatial-to-channel and channel-to-spatial attention mechanisms versus singular attention models on MCC score, Youden's J score and Cohen's kappa score in skin diseases classification.

Model	Avg MCC Score	Youden's J Score	Cohen's Kappa Score
(Spatial → Channel) Attention	**0.91**	**0.88**	**0.91**
Only Channel Attention	0.89	0.86	0.89
Only Spatial Attention	0.90	0.87	0.90
(Channel → Spatial) Attention	0.88	0.85	0.89

TABLE 7. Performance metrics comparison for attention mechanism applied after various layers.

Attention Applied	Accuracy	Precision	Recall	F1-score	Specificity	Kappa Score	Youden's J Score	MCC Score
After Layer 14	0.8996	0.8862	0.8640	0.8723	0.9857	0.8812	0.8497	0.8598
After Layer 15	0.8726	0.8410	0.8233	0.8299	0.9819	0.8492	0.8053	0.8135
After Layer 16	0.8996	0.8826	0.8600	0.8671	0.9857	0.8811	0.8457	0.8556
After Layer 17	**0.9228**	**0.9048**	**0.8942**	**0.8984**	**0.9890**	**0.9086**	**0.8831**	**0.8882**
After Layer 18	0.8953	0.8810	0.8573	0.8591	0.9854	0.8764	0.8427	0.8505
After Layer 14 & 15	0.8846	0.8644	0.8527	0.8506	0.9840	0.8640	0.8367	0.8392
After Layer 15 & 16	0.8958	0.8804	0.8570	0.8649	0.9851	0.8766	0.8421	0.8524
After Layer 16 & 17	0.9035	0.8753	0.8637	0.8679	0.9864	0.8859	0.8501	0.8554
After Layer 17 & 18	0.9035	0.8853	0.8640	0.8716	0.9861	0.8857	0.8502	0.8598
After Layer 14 & 15 & 16	0.8996	0.8720	0.8638	0.8668	0.9858	0.8813	0.8496	0.8534
After Layer 15 & 16 & 17	0.9112	0.8814	0.8746	0.8748	0.9878	0.8952	0.8624	0.8645
After Layer 16 & 17 & 18	0.8996	0.8753	0.8626	0.8662	0.9859	0.8814	0.8485	0.8538

FIGURE 17. Visualizing the influence of features on the proposed model's predictions with LIME for enhanced interpretability.

have achieved marginally higher accuracy rates. Specifically, the study in [25], which employs ResNet-101, demonstrates an accuracy of 99% on the MSID dataset. Additionally, the ensemble approach detailed in [49] achieves an accuracy of 98.7%, slightly edging out our model's performance on the same dataset.

TABLE 8. Comparative analysis of monkeypox detection techniques.

Ref	Dataset	Acc.	Methods
[20]	MSLD	93.39%	Ensemble: Xception, InceptionV3, DenseNet169
[50]	MSID	87.13%	Ensemble: Xception, DenseNet169
[1]	MSID	93.19%	DenseNet-201
[14]	MSID	91.37%	MobileNetV2
[26]	MSID	93%	ViT-B
[51]	MSLD	91.11%	MobileNetV2
[49]	MSID	98.7%	Ensemble: DenseNet-121, ResNet-152V2, ResNet-50, InceptionV3, EfficientNetv2B3
[25]	MSID	99%	ResNet-101
This Study	MSLD	93.33%	Attention-based MobileNetV2
	MSID	98.19%	Attention-based MobileNetV2
	EMSID	92.28%	Attention-based MobileNetV2

These surpassing methods, however, come with inherent trade-offs, particularly in terms of model complexity and computational requirements. ResNet-101, for instance, is a deep convolutional network with over 44.5 M parameters, which significantly increases the computational load and memory requirements. Similarly, the ensemble technique that combines multiple deep learning models, including DenseNet-121, ResNet-152, and EfficientNetv2, further multiplies the computational complexity due to the aggregation of several heavyweight models. These methods are not suitable for the edge devices.

In contrast, our proposed method capitalizes on the lightweight architecture of MobileNetV2, enhanced with attention mechanisms, to offer a more resource-efficient solution. Despite using a considerably lesser number of parameters (3.7 M), our approach achieves an accuracy of 98.19% on the MSID dataset, which is remarkably close to the top-performing methods. This efficiency becomes particularly critical when deploying models on edge devices or in resource-constrained environments, where computational resources and power are limited.

E. EXPLAINABILITY

Explainability in machine learning, especially in high-stakes domains like healthcare, is not a luxury but a necessity. It is crucial for validating model decisions, ensuring reliability, and facilitating trust among practitioners. For this study, we have used two models for the interpretability of the decisions of the deep learning models. We have used LIME [52] and Grad-CAM [53]. LIME and Grad-CAM are prominent interpretability techniques that provide complementary insights into model predictions. Combining these methods can offer a more comprehensive understanding of model behavior. In medical diagnostics, where interpretability is paramount, employing both LIME and Grad-CAM offers a multifaceted view of model decisions. LIME's local, feature-based explanations complement Grad-CAM's visualization of influential regions within the input space, providing both the 'why' and 'where' aspects of a model's prediction. This combination enhances trust in AI-assisted decision-making by ensuring that both local and global features are accounted for and also aids in validating the model's focus on clinically relevant areas of an image. Therefore, the synergy between LIME and Grad-CAM leads to a more transparent, comprehensible, and reliable machine learning model, bolstering its acceptance among healthcare professionals. Next, we describe the working procedure of LIME and Grad-CAM briefly:

1) CONCEPT OF LIME

LIME is premised on the notion that while complex models can be inscrutable, their predictions can be locally approximated by more straightforward, comprehensible models. It generates an interpretable model that is locally faithful to the classifier's behavior, providing insights into the decision-making process for individual predictions. The LIME methodology encompasses several key steps:

1) **Sample Generation**: This involves the creation of a neighborhood dataset around a particular instance x by introducing minor perturbations. This new dataset serves as a basis for understanding the model's behavior near x.
2) **Model Prediction**: The original, complex model is then applied to this neighborhood dataset to predict outcomes for these perturbed samples.
3) **Weight Assignment**: Each sample in the neighborhood is assigned a weight based on its similarity to the original instance x, effectively quantifying its relevance to the instance under investigation.
4) **Surrogate Model Training**: A simpler model, like linear regression, is trained on this weighted dataset. This surrogate model aims to mimic the complex model's behavior in the vicinity of x.
5) **Interpretation**: The final step involves interpreting the surrogate model, particularly its parameters, to infer the complex model's reasoning for the specific instance x.

a: MATHEMATICAL FORMULATION

The core of LIME lies in its mathematical formulation, represented as follows: Let f be the original complex model and g the simpler, interpretable model. Let x' represent a transformed version of x for the interpretable model. LIME aims to find a g that minimizes the locality-aware loss function $\mathcal{L}$, defined as [54]:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z \in Z} \pi_x(z) \left(f(z) - g(x')\right)^2 \tag{28}$$

where:

- Z is the set of perturbed samples.
- $\pi_x(z)$ denotes the proximity weight for each perturbed sample z, calculated using an exponential kernel function:

$$\pi_x(z) = \exp\left(-\frac{||x - z||^2}{2\sigma^2}\right)$$

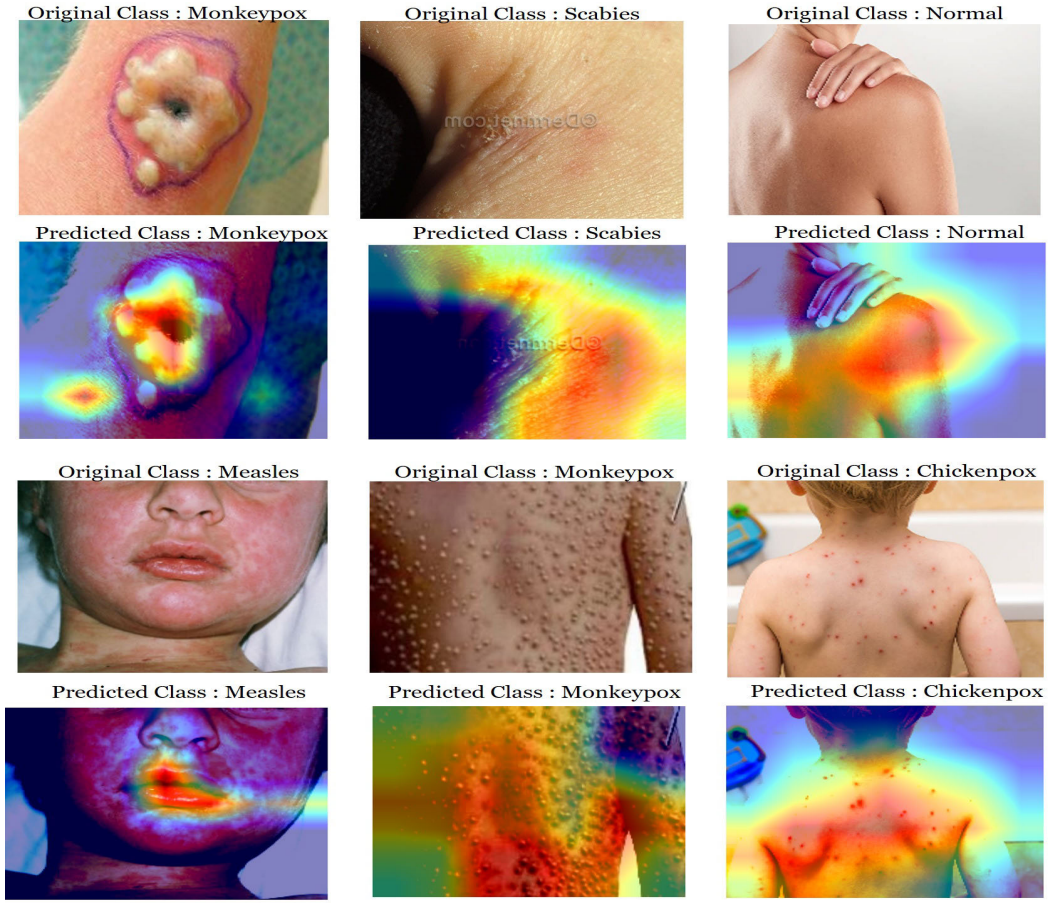

FIGURE 18. Activation map of the proposed model using GradCam.

with σ being a parameter controlling the neighborhood's scope.

Through this optimization process, LIME elucidates the features most influential in the model's prediction for a given instance x. Fig. 17 represents a graphical depiction illustrating how the LIME algorithm identifies and weights the most relevant features for specific predictions. This visualization aids in understanding the local interpretability provided by LIME, highlighting the areas or features in the input data that impact the model's prediction most. As can be seen in the figure, the most important areas are marked in a yellow-colored region.

2) CONCEPT OF GRAD-CAM

Grad-CAM, or Gradient-weighted Class Activation Mapping, leverages the gradients of the target output from a convolutional neural network to produce a heatmap that highlights the important regions for making a prediction. This technique provides visual explanations for the network's decisions, enhancing the model's interpretability. The Grad-CAM algorithm includes the following steps:

1) **Forward Pass**: Conduct a forward pass through the network to obtain the class scores.
2) **Compute Gradients**: Calculate the gradients of the score for the target class with respect to the feature maps of a convolutional layer.
3) **Neuron Importance Weights**: Apply global average pooling to the gradients to derive the neuron importance weights.
4) **Generate Grad-CAM Heatmap**: Produce the heatmap by weighting the feature maps with the neuron importance weights, summing them, and applying the ReLU function.

a: MATHEMATICAL FORMULATION

The mathematical formulation for Grad-CAM is as follows. Given A^k as the k-th feature map of a convolutional layer, and y^c as the score for the target class c before the softmax layer, the neuron importance weight α_k^c is calculated by [55]:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{29}$$

where Z is the total number of pixels in the feature map A^k. The Grad-CAM L^c for class c is then computed as [55]:

$$L^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{30}$$

The ReLU function is applied to the weighted combination of feature maps to ensure that only the features that positively influence the score of the target class contribute to the heatmap. This application of ReLU effectively filters out the negative values in the feature maps, highlighting areas of importance that positively impact the class prediction. Fig.18 shows the activation map of some instances, demonstrating how different regions in the input images are activated for the target class. These activation maps visually represent the areas and features in the image most influential in determining the model's classification decision, providing insights into the model's focus and decision-making process.

VII. CONCLUSION

In conclusion, our study delves into leveraging artificial intelligence for early detection of monkeypox, a public health issue that has become increasingly significant post-COVID-19. The complexity of the medical diagnosis process, particularly regarding early-stage infections, presents a notable challenge. Our study focuses on the MobileNetV2 model, characterized by its lightweight structure and low parameter count, making it suitable for operation on mobile phones and other intelligent edge devices. We have improved accuracy by incorporating spatial and channel attention mechanisms, effectively distinguishing monkeypox from similar dermatological conditions. Compared with different established architectures like ResNet-152, VGG-19, GoogLeNet, etc., our model's performance demonstrates superior accuracy and efficiency, making it particularly valuable in resource-constrained settings. We further enhance our model's applicability in medical settings by integrating Grad-CAM and LIME, ensuring the interpretability and reliability of the AI-driven recommendations for healthcare professionals. Extensive evaluation using metrics such as Cohen's Kappa, MCC, and Youden's J Index confirms the model's proficiency, with it achieving impressive accuracies of 92.28% on the EMSID dataset and 98.19% on the MSID dataset. These results underscore the proposed model's potential in aiding monkeypox diagnosis, contributing significantly to public health efforts.

REFERENCES

[1] D. Bala, M. S. Hossain, M. A. Hossain, M. I. Abdullah, M. M. Rahman, B. Manavalan, N. Gu, M. S. Islam, and Z. Huang, "MonkeyNet: A robust deep convolutional neural network for monkeypox disease detection and classification," *Neural Netw.*, vol. 161, pp. 757–775, Apr. 2023.

[2] M. Dwivedi, R. G. Tiwari, and N. Ujjwal, "Deep learning methods for early detection of monkeypox skin lesion," in *Proc. 8th Int. Conf. Signal Process. Commun. (ICSC)*, Dec. 2022, pp. 343–348.

[3] A. S. Jaradat, R. E. Al Mamlook, N. Almakayeel, N. Alharbe, A. S. Almuflih, A. Nasayreh, H. Gharaibeh, M. Gharaibeh, A. Gharaibeh, and H. Bzizi, "Automated monkeypox skin lesion detection using deep learning and transfer learning techniques," *Int. J. Environ. Res. Public Health*, vol. 20, no. 5, p. 4422, Mar. 2023.

[4] (2023). *Multi-Country Outbreak of Mpox*. Accessed: Nov. 23, 2023. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20230814_mpox_external-sitrep-27.pdf

[5] F. Yasmin, Md. M. Hassan, M. Hasan, S. Zaman, C. Kaushal, W. El-Shafai, and N. F. Soliman, "PoxNet22: A fine-tuned model for the classification of monkeypox disease using transfer learning," *IEEE Access*, vol. 11, pp. 24053–24076, 2023.

[6] M. Gain and R. Debnath, "A novel unbiased deep learning approach (DL-Net) in feature space for converting gray to color image," *IEEE Access*, vol. 11, pp. 78918–78933, 2023.

[7] M. Gain, M. A. Rahman, R. Debnath, M. M. Alnfiai, A. Sheikh, M. Masud, and A. K. Bairagi, "An improved encoder–decoder CNN with region-based filtering for vibrant colorization," *Comput. Syst. Sci. Eng.*, vol. 46, no. 1, pp. 1059–1077, 2023.

[8] M. Gain, A. D. Raha, A. Adhikary, and C. S. Hong, "Transfer learning based face mask detection using deep neural networks," in *Proc. Korea Comput. Congr.* South Korea: Korea Information Science Society, 2023, pp. 1377–1379.

[9] S.-Y. Lu, S.-H. Wang, and Y.-D. Zhang, "SAFNet: A deep spatial attention network with classifier fusion for breast cancer detection," *Comput. Biol. Med.*, vol. 148, Sep. 2022, Art. no. 105812.

[10] A. Sutradhar, M. Al Rafi, P. Ghosh, F. J. M. Shamrat, M. Moniruzzaman, K. Ahmed, A. Azad, F. M. Bui, L. Chen, and M. A. Moni, "An intelligent thyroid diagnosis system utilising multiple ensemble and explainable algorithms with medical supported attributes," *IEEE Trans. Artif. Intell.*, 2023, doi: 10.1109/TAI.2023.3327981.

[11] A. Tilve, S. Nayak, S. Vernekar, D. Turi, P. R. Shetgaonkar, and S. Aswale, "Pneumonia detection using deep learning approaches," in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng. (ic-ETITE)*, Feb. 2020, pp. 1–8.

[12] M. Alazab, A. Awajan, A. Mesleh, A. Abraham, V. Jatana, and S. Alhyari, "COVID-19 prediction and detection using deep learning," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 12, pp. 168–181, Jan. 2020.

[13] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," *Healthcare*, vol. 10, no. 3, p. 541, Mar. 2022.

[14] S. Nafisa Ali, M. Tazuddin Ahmed, J. Paul, T. Jahan, S. M. S. Sani, N. Noor, and T. Hasan, "Monkeypox skin lesion detection using deep learning models: A feasibility study," 2022, *arXiv:2207.03342*.

[15] S. Majumder and M. J. Deen, "Smartphone sensors for health monitoring and diagnosis," *Sensors*, vol. 19, no. 9, p. 2164, May 2019.

[16] Y. Chen, S. Biookaghazadeh, and M. Zhao, "Exploring the capabilities of mobile devices in supporting deep learning," in *Proc. 4th ACM/IEEE Symp. Edge Comput.*, Nov. 2019, pp. 127–138.

[17] A. Deb Raha, M. Shirajum Munir, A. Adhikary, Y. Qiao, and C. S. Hong, "Generative AI-driven semantic communication framework for NextG wireless network," 2023, *arXiv:2310.09021*.

[18] T. Kujani, S. Alex David, T. Sathya, P. Arivubrakan, and S. Shanmuga Priya, "Efficient brain tumor detection using VGG-16 and ResNet50 transfer learning models," in *Proc. Int. Conf. Soft Comput. Secur. Appl.* Singapore: Springer, 2023, pp. 455–467.

[19] A. S. Sathwik, B. Naseeba, J. C. Kiran, K. Lokesh, V. S. D. Ch, and N. P. Challa, "Early detection of monkeypox skin disease using patch based DL model and transfer learning techniques," *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 9, pp. 1–9, Nov. 2023.

[20] R. Pramanik, B. Banerjee, G. Efimenko, D. Kaplun, and R. Sarkar, "Monkeypox detection from skin lesion images using an amalgamation of CNN models aided with beta function-based normalization scheme," *PLoS ONE*, vol. 18, no. 4, Apr. 2023, Art. no. e0281815.

[21] A. Deb Raha, A. Adhikary, M. S. Munir, Y. Qiao, and C. S. Hong, "Segment anything model aided beam prediction for the millimeter wave communication," in *Proc. 24st Asia–Pacific Netw. Operations Manage. Symp. (APNOMS)*, Sep. 2023, pp. 113–118.

[22] A. Shah, "Monkeypox skin lesion classification using transfer learning approach," in *Proc. IEEE Bombay Sect. Signature Conf. (IBSSC)*, Dec. 2022, pp. 1–5.

[23] M. Manjurul Ahsan, M. Shahin Ali, M. Mehedi Hassan, T. Abu Abdullah, K. Datta Gupta, U. Bagci, C. Kaushal, and N. F. Soliman, "Monkeypox diagnosis with interpretable deep learning," *IEEE Access*, vol. 11, pp. 81965–81980, 2023.

[24] F. B. Demir, M. Baygin, I. Tuncer, P. D. Barua, S. Dogan, T. Tuncer, C. P. Ooi, E. J. Ciaccio, and U. R. Acharya, "MNPDenseNet: Automated monkeypox detection using multiple nested patch division and pretrained DenseNet201," *Multimedia Tools Appl.*, pp. 1–23, 2024, doi: 10.1007/s11042-024-18416-4.

[25] M. M. Ahsan, M. R. Uddin, M. S. Ali, M. K. Islam, M. Farjana, A. N. Sakib, K. A. Momin, and S. A. Luna, "Deep transfer learning approaches for monkeypox disease diagnosis," *Expert Syst. Appl.*, vol. 216, Apr. 2023, Art. no. 119483.

[26] M. A. Arshed, H. A. Rehman, S. Ahmed, C. Dewi, and H. J. Christanto, "A 16×16 patch-based deep learning model for the early prognosis of monkeypox from skin color images," *Computation*, vol. 12, no. 2, p. 33, 2024.

[27] M. G. Campana, M. Colussi, F. Delmastro, S. Mascetti, and E. Pagani, "A transfer learning and explainable solution to detect mpox from smartphones images," *Pervasive Mobile Comput.*, vol. 98, Feb. 2024, Art. no. 101874.
[28] (2023). *Dermnet*. Accessed: Sep. 2023. [Online]. Available: http://www.dermnet.com/
[29] S. Dey, *Hands-On Image Processing With Python: Expert Techniques for Advanced Image Analysis and Effective Interpretation of Image Data*. Birmingham, U.K.: Packt Publishing, 2018.
[30] S. Gopal Krishna Patro and K. Kumar Sahu, "Normalization: A preprocessing stage," 2015, *arXiv:1503.06462*.
[31] S. M. Islam and R. Debnath, "A comparative evaluation of feature extraction and similarity measurement methods for content-based image retrieval," *Int. J. Image, Graph. Signal Process.*, vol. 12, no. 6, pp. 19–32, 2020.
[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.
[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[37] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.
[38] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6688–6697.
[39] H. Lee, J. Park, and J. Y. Hwang, "Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 7, pp. 1344–1353, Jul. 2020.
[40] Y.-H. Jin et al., "A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version)," *Mil. Med. Res.*, vol. 7, no. 1, p. 4, 2020.
[41] D. Alghazzawi, O. Rabie, O. Bamasaq, A. Albeshri, and M. Z. Asghar, "Sensor-based human activity recognition in smart homes using depthwise separable convolutions," *Hum.-Centric Comput. Inf. Sci.*, vol. 12, p. 50, Jan. 2022.
[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[43] C. Peng and J. Ma, "Semantic segmentation using stride spatial pyramid pooling and dual attention decoder," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107498.
[44] Y. Qu, M. Xia, and Y. Zhang, "Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow," *Comput. Geosci.*, vol. 157, Dec. 2021, Art. no. 104940.
[45] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," *Neural Comput. Appl.*, vol. 35, no. 23, pp. 17095–17112, Aug. 2023.
[46] I. K. Nti, O. N. Boateng, A. F. Adekoya, B. A. Weyori, and H. P. Adjei, "Predicting diabetes using Cohen's Kappa blending ensemble learning," *Int. J. Electron. Healthcare*, vol. 13, no. 1, p. 57, 2023.
[47] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Mining*, vol. 16, no. 1, pp. 1–23, Feb. 2023.
[48] Y. Liu, B. Lawson, X. Huang, B. Broom, and J. Weinstein, "RETRACTED: Prediction of ovarian cancer response to therapy based on deep learning analysis of histopathology images," *Cancers*, vol. 15, no. 16, p. 4044, Aug. 2023.
[49] R. Haque, A. Sultana, and P. Haque, "Ensemble of fine-tuned deep learning models for monkeypox detection: A comparative study," in *Proc. 4th Int. Conf. Emerg. Technol. (INCET)*, May 2023, pp. 1–8.
[50] C. Sitaula and T. B. Shahi, "Monkeypox virus detection using pre-trained deep learning-based approaches," *J. Med. Syst.*, vol. 46, no. 11, p. 78, Oct. 2022.
[51] V. H. Sahin, I. Oztel, and G. Yolcu Oztel, "Human monkeypox classification from skin lesion images with deep pre-trained network using mobile application," *J. Med. Syst.*, vol. 46, no. 11, p. 79, Oct. 2022.
[52] S. Ahmad, M. G. R. Alam, J. Uddin, M. R. Bhuiyan, and T. S. Apon, "Machine learning based stream selection of secondary school students in Bangladesh," *Indonesian J. Electr. Eng. Informat. (IJEEI)*, vol. 11, no. 1, pp. 105–118, Feb. 2023.
[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[54] A. Aljadani, B. Alharthi, M. A. Farsi, H. M. Balaha, M. Badawy, and M. A. Elhosseini, "Mathematical modeling and analysis of credit scoring using the LIME explainer: A comprehensive approach," *Mathematics*, vol. 11, no. 19, p. 4055, Sep. 2023.
[55] A. Abhishek, R. K. Jha, R. Sinha, and K. Jha, "Automated detection and classification of leukemia on a subject-independent test dataset using deep transfer learning supported by grad-CAM visualization," *Biomed. Signal Process. Control*, vol. 83, May 2023, Art. no. 104722.

AVI DEB RAHA (Graduate Student Member, IEEE) received the B.S. degree in computer science from Khulna University, Bangladesh, in 2020. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Kyung Hee University, South Korea. His research interests are currently focused on deep learning, generative AI, semantic communication, and integrated sensing and communication.

MRITYUNJOY GAIN received the B.S. degree in computer science from Khulna University, Bangladesh, in 2021. He is currently pursuing the combined M.S. and Ph.D. degrees in computer science with Kyung Hee University, South Korea. He is also with the Networking Intelligence Laboratory. His research interests include image processing, deep learning, meta-learning, deep feature engineering, explainable artificial intelligence, and pattern recognition and their applications.

RAMESWAR DEBNATH (Senior Member, IEEE) received the bachelor's degree (Hons.) in computer science and engineering from Khulna University, Bangladesh, in 1997, and the Master of Engineering degree in communication and systems and the Ph.D. degree from The University of Electro-Communications, Tokyo, in 2002 and 2005, respectively. He is a Professor of the Computer Science and Engineering Discipline, Khulna University. He was a Postdoctoral Researcher with the Department of Informatics, The University of Electro-Communications, and Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology, Tsukuba, under JSPS Fellowship, from 2008 to 2010. He was the Head of the Computer Science and Engineering Discipline, Khulna University, from 2012 to 2015. His research interests include image data analysis, deep learning, bioinformatics, support vector machine, artificial neural networks, statistical pattern recognition, and medical image processing. He received the Japanese Government Scholarship for the Ph.D. study. He was the Organizing Chair of the 16th International Conference on Computer and Information Technology (ICCIT), in 2014.

APURBA ADHIKARY (Student Member, IEEE) received the B.Sc. and M.Sc. (Eng.) degrees in electronics and communication engineering from Khulna University, Khulna, Bangladesh, in 2014 and 2017, respectively. He is the Ph.D. Researcher with the Department of Computer Science and Engineering, Kyung Hee University (KHU), South Korea. He has been an Assistant Professor with the Information and Communication Engineering Department, Noakhali Science and Technology University (NSTU), Noakhali, Bangladesh, since January 2020. In addition, he was a Lecturer with the Information and Communication Engineering Department, NSTU, from January 2018 to January 2020. His current research interests focused on integrated sensing and communication, holographic MIMO, cell-free MIMO, intelligent networking resource management, artificial intelligence, and machine learning.

YU QIAO (Graduate Student Member, IEEE) received the B.E. degree in the Internet of Things engineering and the M.E. degree in computer science and technology from Nanjing University of Information Science and Technology (NUIST), Nanjing, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Artificial Intelligence, Kyung Hee University (KHU), South Korea. Before the Ph.D. study, he was a Camera Software Engineer with Spreadtrum Communications (UNISOC), Shanghai, China, from 2019 to 2022. His research interests include machine learning, federated learning, self-supervised learning, and distributed edge intelligence.

MD. MEHEDI HASSAN (Member, IEEE) received the B.Sc. (Eng.) degree in computer science and engineering, in 2022. He is currently pursuing the M.Sc. (Eng.) degree with Khulna University, Bangladesh. He is a dedicated young researcher. His remarkable aptitude for research has propelled him to excel in biomedical engineering, data science, and expert systems, earning him recognition as a respected leader in these fields. He is the founder and a CEO of the Virtual BD IT Firm and the Laboratory Head of the VRD Research Laboratory, Bangladesh. His impactful work, published by esteemed publishers such as Nature, Elsevier, IEEE, Springer, and more, those are significantly contributed to the advancement of knowledge in his field. With over three filed patents, three of which have been granted. He is not only an innovative thinker but also a practical problem solver. His research interests encompass a broad spectrum, ranging from human brain imaging, neuroscience, machine learning, and artificial intelligence to software engineering. Driven by his notable accomplishments and promising potential, he remains dedicated to leveraging cutting-edge scientific research to enhance human health and well-being. He serves as a reviewer for prestigious journals, further underscoring his influence in the scientific community.

ANUPAM KUMAR BAIRAGI (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science and engineering from Khulna University (KU), Bangladesh, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea. He is a Professor of the Discipline of Computer Science and Engineering, KU. His research interests include wireless resource management in 5G, cooperative communication, game theory, health informatics, the IIoT, and UAV communication.

SHEIKH MOHAMMED SHARIFUL ISLAM is a Physician Scientist with the Institute for Physical Activity and Nutrition, Deakin University. He is an Honorary Senior Lecturer with Sydney Medical School, University of Sydney; and an Honorary Senior Fellow with George Institute for Global Health, UNSW. Previously, he led the Diabetes Research Program, Center for Control of Chronic Diseases, ICDDR,B. He has working experience with the United Nations Development Program, World Health Organization, and German Technical Cooperation, as a Resident Cardiologist in Bangladesh. He has more than 100,000 citations to his scholar profile.