Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 19 No. 16 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v19i16.44461

PAPER

An Adaptable Model for Medical Image Classification Using the Streamlined Attention Mechanism

Dakshayani Himabindu Damineni[1](✉), Praveen Kumar Sekharamantry[2], Rajakoti Badugu[2]

[1]Department of CSE, Research Scholar, GST, GITAM University, Andhra Pradesh, India

[2]Department of CSE, Faculty of CSE, GST, GITAM University, Andhra Pradesh, India

dakshayani.himabindu@gmail.com

ABSTRACT

The resurgence of deep learning has improved computer vision by increasing its applicability and scalability for challenges in the real world. Specifically, utilizing attention in computer vision tasks has improved the performance of models to a superior level. Today we need medical diagnostic tools to tackle the immediate needs of the population suffering from Cancer and COVID-19. Thus, an end-to-end screening is tedious and needs validation expertise. But, with the current deep learning methods, it is quite possible to provide a diagnostic tool that can assist doctors and patients with their immediate needs. So, to tackle these real-world problems, we have proposed a method that is implied on 3 diverse standard datasets in the field of medical imagery, which are Skin Lesions, Brain tumors, and COVID-19 classification. To justify the model's performance, the authors have experimented on 5 diverse data sets ranging from binary class to multi-class. The experimentation has shown that the proposed "streamlined-attention module" is not only capable of producing superior performance in fine-grained visual recognition but also bio-medical imagery. To further justify, we have illustrated Grad-Cam heat map visualizations to the model and show that it can extract the detailed features with proportionate attention. Our results have proved that our methods excel in their performance compared to that of existing methods. It has achieved state-of-the-art accuracy scores on COVID-19 and HAM10000 datasets with a well-guided explainable result. This work represents a significant advancement in the field of medical image processing with clear results, and the authors anticipate that the suggested method will prove to be a useful tool for medical professionals in the detection and identification of diseases like Covid-19 and cancer. This paper provided the best accuracy for COVID-19 multiclass **(94.75 ± 1.07)** and HAM10000 **(94.31 ± 0.91)**.

KEYWORDS

medical image classification, visual attention, deep learning, COVID-19, skin cancer, brain tumor

1 INTRODUCTION

The paradigm shift from extracting features from the data to extracting feature maps from the raw inputs via deep neural networks [1] is considered evolutionary.

Damineni, D.H., Sekharamantry, P.K., Badugu, R. (2023). An Adaptable Model for Medical Image Classification Using the Streamlined Attention Mechanism. *International Journal of Online and Biomedical Engineering (iJOE)*, 19(16), pp. 93–110. https://doi.org/10.3991/ijoe.v19i16.44461

Article submitted 2023-08-18. Revision uploaded 2023-09-23. Final acceptance 2023-09-23.

Mainly, the evolution of convolutional neural networks [2–4] has produced the most reliable results for computer vision tasks. This helped in deriving efficient feature maps, which involved better patterns that the machine could easily interpret.

The tendency of machines to understand such feature maps is mainly due to certain feature embeddings formed at intermediate layers of a specific architecture. This not only helps in understanding generic data, which involves multiple classes but is considered equally important in the field of medical imagery for classifying the data. This data mainly involves visual information such as (Computed Tomography) CT scan, (Chest X-Ray) CXR, (Magnetic Resonance Imaging) MRI, and dermatoscopic images, which are collected via various sources from the medical industry. Each image input serves a specific purpose to solve a certain kind of problem under classification. The details concerning each dataset have been elaborated in Section 3, along with their purpose.

The methods by which the data are classified in this specific space are considered automatable, scalable, and can be comfortably deployed in a specific setting without much effort. The scope of image analysis in the medical field is considered to be highly broad [5–8]. Hence, the methods developed for such tasks must be robust enough to predict effective outputs quickly. As these methods are developed with a notion of deployment, they must be developed with conscientiousness and integrity.

The necessity of such methods in the field is momentous. Manual recognition of medical issues via CT-Scan images to identify COVID-19, MRI images to identify brain tumors, or dermatoscopic images to identify skin lesions is time-consuming and tedious. Identifying such issues mainly involving easily transmissible viruses must be rapid. Hence, to address such problems, researchers have developed various computer-aided diagnostic (CAD) techniques [9]. These techniques are based on feature maps acquired from deep convolutional neural networks.

Hence, to address the problem of image classification on certain datasets i.e.

1. Skin Cancer Classification:
   a) HAM10000 Dataset [10] (Multiclass Classification Problem)
2. Covid-19 Classification:
   a) SARS-COV-2 CT-Scan Dataset [11] (Binary Classification Problem)
   b) Chest X-Ray COVID-Classifier [12] (Multiclass Classification Problem)
3. Brain Tumor Detection:
   a) (Binary Classification Problem) [13]
   b) (Multiclass Classification Problem)

The proliferation of deep learning [12] worldwide has been remarkable in its ability to provide a more profound understanding of convolutional neural networks (CNN) [13–14]. We have developed an attention-based mechanism. This is inspired by our previous work [15], in which we tested our model on standard classification data to address the problem of image classification and fine-grained visual recognition. After gaining significant results with state-of-the-art performance on one of the datasets, we found enough scope to broaden the perspective toward the medical field as well. The initial model was created by Bahdanau et al. [16] in the field of natural language processing to address the neural machine translation tasks inspired by attention-based method. This was concerning the features which were focusing on valuable information on a specific feature map. Hence, we tried to embed this novel concept into computer vision with appropriate ailments.

The advent of deeply layered convolutional neural networks [8, 17] has enabled us to extract more meaningful features and patterns, resulting in breakthrough

performances in computer vision tasks such as image classification, segmentation, recognition, and detection. Several techniques are employed in neural networks for computer vision tasks, one of the most influential being attention-based modeling. An attention mechanism was first introduced [1, 18] to improve feature-based performance by harvesting important information from a neural network in the context of machine translation tasks. Attention mechanisms are vital components of the human visual system, as they enable us to focus on the important parts of a scene instead of the entire thing, and they are just as effective in computer vision tasks. A variety of efficient approaches have been classified and structured for attention extraction. Neural machine translation has achieved remarkable results in transduction tasks [6, 19, 20], and attention-based mechanisms are beneficial in vision tasks, particularly in image captioning. This has highlighted the correlation between visual features and their corresponding text generation. These successful approaches include global and local attention techniques, which employ soft and hard attention [21]. Global attention is accomplished through the "soft attention approach," which determines the aligned weights and the features extracted from the entire data, consequently leading to considerable computational expense. Soft attention is relatively low performing due to its tendency to generate attentive weighted sum averages from the entire image patch. On the other hand, local attention combines soft and hard attention, with hard attention focusing on sub-patches of the image.

Global attention is considered computationally expensive as it tends to align the weights of the whole image and gain an attentive weighted sum average from it. A local attention approach is considered relatively efficient since it collects relevant information from every sub-patch of a larger area. The application of these approaches is crucial. Hence, to develop such an attention-based mechanism, two mechanisms exist addressing the required partials of the image. They are Channel and Spatial attention. The details regarding these mechanisms will be provided in Section 3.

In this paper, we have applied medical imagery-based data for classifying images using a streamlined attention-based module. Efficient channel and spatial attention are integrated with a second-order pooling block, a mathematical operation for gaining the outer product of two matrices, in this case, two spatial feature maps.

Even though this model was initially built for classifying standard computer vision data, for bringing to light the work of hybrid attention networks, this model generalizes well on medical data by capturing the infection efficiently on each medical image. For example, the ground glass opacity on lung images has been paid equal attention as observed via GradCAM visualizations, considering COVID-19 data results have been addressed and provided appropriate reasoning in Section 4.

Applying the attention mechanism in architecture by inducing an efficient module into the existing architecture is unique in our procedure. As this module involves a mathematical operation that cuts down the number of parameters concentrating on required information for classification purposes, this can be considered to be efficient for a reason.

2 CONTRIBUTION

The following list includes this paper's significant contributions:

a) For medical data, we have put forth a novel attention-based module that combines channel and spatial attention.

b) The proposed model shows significant performance even on cross-domain i.e., the performance does not vary even if the model was trained on any other medical imagery and provides significant results for both standard computer vision and medical imagery.
c) The experimental results were consistent and reliable without any augmentation applied to the model. So, without any augmentations, we have achieved state-of-the-art performance on HAM10000 and COVID-19 datasets (both binary and multiclass).
d) For providing interpretable results, the authors have justified the GradCAM visualizations. This gives a visual perception of how the model can acquire rich features. This analysis has been performed on all the datasets and details how the proposed model is better than the existing literature.

Further, this paper is elaborated into multiple sections where Section 3 explains the existing work and its drawbacks. Section 4 elaborates on the method proposed and Section 5 discusses the corresponding results. Sections 6 and 7 are the work's future scope and conclusion.

3 RELATED WORK

The HAM10000 dataset, also referred to as Human Against Machine, was proposed by Philipp Tschandl et al. [10] and comprised 10,000 training images considered dermatoscopic images. The primary aim of creating this dataset was to address the predicament faced when diagnosing pigmented skin lesions. A significant issue with this dataset is its limited size and lack of diversity, which this work primarily focuses on by balancing the size of melanoma and nevi-based data. This dataset is considered one of the most commonly used datasets for skin lesions. Further information regarding the dataset can be found in Section 3.

According to Kyle Young et al. [18], models were tested through a Bayesian hyper-parameter search for specified metrics using the HAM10000 dataset. They even investigated the dependability of GradCAM and KernelSHAP through a few sanity-check experiments to bridge the gap between the precision and interpretability of the models. Ardan Adi et al. [18] presented a CNN-based skin lesions dataset classification. Hsin-Wei et al. [19] concentrated on binary and multiclass classification problems using the HAM10000 and KCGMH datasets. They combined EfficientNet and DenseNet on preprocessed images including color jitter, vertical flip, and horizontal flip. Muhammad Attique Khan et al. [22] proposed a deep learning-based framework utilizing Mask-RCNN and Transfer Learning on ISBI2016, ISI2017, and HAM10000 datasets. They utilized a novel entropy-controlled least square Support Vector Machine (SVM) for optimization before the classification phase and applied the Extreme Learning Machine (ELM) methodology.

Farhat Afza et al. [23] described how an extreme learning machine was utilized to produce a new method combining deep learning features with an extreme learning machine for image acquisition and contrast enhancement. Transfer learning was employed to extricate deep learning features, while optimization was used to maximize entropy and mutual information, and a modified canonical correlation-based method was utilized to fuse chosen features. The classification process was ultimately accomplished with the help of an extreme learning machine-based approach, demonstrating the computational efficiency of the model. Ren et al. [20] proposed an attention fusion mechanism involving spatial and channel information to segment skin images by extracting information from suggested intermediate channels.

Soares et al. constructed a SARS-CoV-2 CT scan image dataset [11], collected from a hospital in Sao Paulo, Brazil. Furthermore, this document proposes an explicable Deep Learning technique (xDNN) with VGGNet as an encoder, solving the problem of classifying the proposed dataset. It applies a distance-based approach in classifying the presented data, calculating the distance between data points to allot them to particular classes with a certain set of defined metrics. Subsequently, it compares its findings to conventional machine learning algorithms like SVM. Additionally, Khuzani et al. [12] proffered a chest X-Ray image-based small-scale COVID-19 dataset, employing a dimensionality reduction approach to generate salient features from the chest x-ray images. Moreover, Rahimzadeh et al. [24] have proposed a novel dataset comprising CT scan images based on COVID-19 infection, introducing an architecture involving a pyramid network coupled with ResNet50V2.

The deep transfer learning algorithm proposed by Panwar et al. [25] has conducted Grad-CAM analysis for colored visualization of the infections to accelerate predictions. Moreover, Banerjee et al. [26] proposed a COFE-Net dubbed COVID Fuzzy Ensemble Network, using Choquet fuzzy integral to apply ensemble learning to three major architectures: Inception V3, Inception ResNet V2, and DenseNet 201. DT-BiLTCN architecture consists of bidirectional long short-term memory network is proposed by Raza A. et al. [27].

Rehman et al. [28] advanced novel transfer and deep learning methods for classification and automated brain tumor detection, testing them on the Figshare dataset and exploring transfer learning by experimenting with three architectures, i.e., AlexNet, GoogleNet, and VGGNet. In addition, Kang et al. [29] performed ensemble learning on three pre-trained CNN models. Alanazi et al. [30] proposed a 22-layer CNN developed from the ground up to examine the performance scale of MRI images. Then, they reused the weights and carried out a transfer learning on the same data, resulting in significant results in diagnosing brain tumors. Furthermore, multiple works are exhibiting the implementation of a transfer learning approach by implementing tumor and non-tumor-based data, such as [31].

The disadvantages of the existing works are as follows:

a) Skin lesions caused by melanocytosis are often in uneven shades of black, brown, and tan with white, grey, red, pink, or blue areas, which generally aid in skin cancer recognition. If the pre-processing step applies color jitter [19] as augmentation, then the data loses its capacity to be identified as infected skin or non-infected.

b) In this study [24], 48,260 CT scan images were obtained from 282 normal individuals, and 15,589 CT scan images were gathered from 95 patients with COVID-19, leading to a disparity in the normal:infected ratio, thus causing inadequate model generalization.

c) Most of the works classifying brain tumors are either utilizing generic architectures to conduct transfer learning or ensemble learning, with no specified proposed architecture that involves attention as its architectural module. Therefore, to offset these limitations, we desire to reduce them by testing our model on numerous standard bio-medical data sets and demonstrating that the proposed model is comparatively more proficient in multiple aspects.

4 MATERIALS AND METHODOLOGY

In this section, we describe the materials and methodology used in our study to evaluate the proposed streamlined-attention module. Firstly, we provide an

overview of the dataset used in our experiments (Section 4.1), followed by a description of the proposed method (Section 4.2) and the evaluation metrics used.

Following the introductory text, the subheadings "4.1 Dataset Description" and "4.2 Method" can be introduced. Specifically, under "4.1 Dataset Description," you could provide a description of the dataset used in the study, including its size, characteristics, and any pre-processing steps that were applied to the data. Under "4.2 Method," you could provide a detailed explanation of the proposed streamlined-attention module, including its architecture, training methodology, and the significance of second-order-pooling.

4.1 Dataset description

HAM10000. The HAM10000 dataset [10] involves 10015 dermatoscopic images released via the ISIC archive for educational and academic research-based purposes. Actinic Keratoses (AKEIC), Basal Cell Carcinoma (BCC), Benign Keratosis (BKL), Dermatofibroma (DF), Melanocytic Nevi (NV), Melanoma (MEL), and Vascular Skin Lesions are the seven classes in this dataset (VASC). The pigmented skin lesions listed above pertain to seven distinct types of cancer. There exist a few sets of clinical methods that can help in diagnosing skin cancer. These results are often considered to be non-reproducible and need a great amount of knowledge in this field to differentiate between data.

COVID-19. The COVID-19 disease is caused by the SARS-CoV2 virus and has been widely spread worldwide, impacting thousands of people. Therefore, COVID-19 was deemed a pandemic by the World Health Organization (WHO). The outbreak of COVID-19 happened in Wuhan, China, in December 2019. The analysis by scientists revealed high rates of ground-glass opacities in the CTs of the infected people [32] [33].

a) Binary Classification Data – This dataset includes 2482 CT scan images from which 1252 images are derived from infected patients and the remaining 1230 images are considered to be derived from non-infected patients. [11]
b) Multiclass Classification Data – This dataset comprises 420 2-D Posteroanterior (P.A.) chest view X-ray images that have been further divided into 3 classes. There are three of them: COVID-19, Pneumonia, and Normal (140 photos each) (140 images)

Brain tumor. A human brain cell's unregulated, unnatural, and abnormal growth is referred to as a brain tumor. There are two forms of brain tumors. There are glioma tumors and non-glioma tumors. Glioma tumors are considered to be the most commonly occurring brain tumors, which grow from glial cells. Tumors that form in the brain from cells other than glial cells are referred to as non-glioma tumors [34].

a) Binary Classification Data – This data is collected from the Kaggle website and the appropriate link has been provided [13]. An MRI image of 155 tumors and 98 non-tumors is included in this dataset.
b) Multiclass Classification Data – Kaggle contains 4 classes of MRI images, i.e., glioma tumor, meningioma tumor, no tumor, and pituitary tumor, each containing 826, 822, 395, and 827 images [14].

4.2 Method

This method of applying attention to medical data is considered very valuable as each kind of visual information, i.e., a CT scan, Chest X-Ray, MRI, or dermatoscopic image, is considered to have a few sets of attentive regions concentrated on the specific infection. These infections need to be carefully monitored and captured. To diagnose in such a way, we require attention to our architecture. The relevant reasoning for why one architecture needs attention, especially for such medical data has been discussed in successive sections in detail. With performance in mind, the ResNet [3] module was used as the encoder for the proposed design. This module also goes beyond previous work. In the model architecture, each layer of a typical ResNet module is followed by an additional block of simplified attention modules. In the proposed module, the global average pooling layer, which comes after the fully linked layer, is followed by a second-order pooling algorithm. The following paragraphs provide details about second-order pooling. The suggested module's detailed structure is depicted in detail in Figure 1.

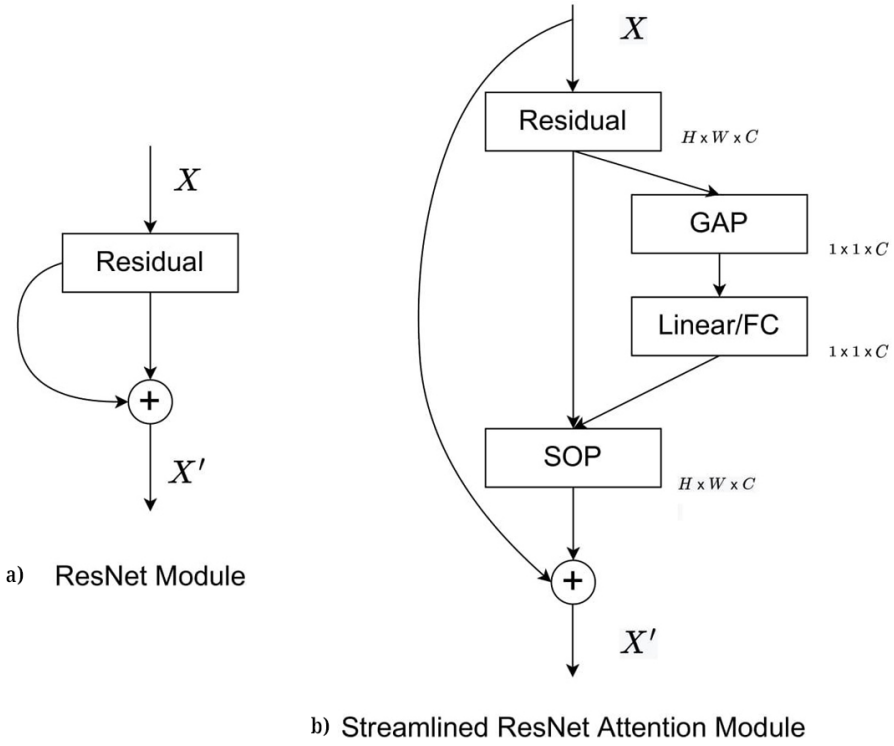

Fig. 1. a) Traditional ResNet residual module, b) SE ResNet attention module; where GAP abbreviates to global average pooling and SOP for second-order pooling. In addition, the sub-components of the figure should be defined explicitly. For example, the Traditional ResNet residual module could be defined as a convolutional neural network architecture that uses residual connections to improve model performance, while the SE ResNet attention module could be defined as a variant of the ResNet architecture that uses channel-wise attention to improve feature extraction

Channel and spatial attention. The proposed simplified attention block module has taken the mechanisms [35–37] into consideration, as these proposed works significantly highlight how effective channel attention might be shown to improve. According to Sanghyun Woo's proposal in [35], the block/combination of acquired feature maps, which are regarded to be adequate to provide long-range feature

dependencies, should also be able to recognize channel-wise dependencies. With fewer additional layers, this mostly aids in delivering high attention. In addition to contemplating "what" should be the subject of a channel attention block, this incorporates the use of the features learned via cross-channel interactions. Following that research, [36] showed similar results by adhering to the "dimensionality reduction" standard with the addition of a kernel that finally took into account the initialization of the neurons. In [37], they have focused on channel attention by considering the appropriate technique of constricting and exciting the feature maps after a certain point. The suggested attention module shows the second-order pooling and global average pooling described previously in Section 4.2. Due to its propensity to capture attention with fewer effective layers, this technique is considered thorough.

Figure 2 demonstrates the outcomes of passing a block of feature maps via the specified channel attention module. The feature maps emphasize the long-range dependencies identified by the global average pooling layer using feature maps of the preceding layers. To bring awareness to a model, global average pooling is vital. As a result, the H, W, and C feature dimensions are condensed into a feature map that utilizes convolution and has a significant amount of useful data in the C channels. A second objective is accomplished by training a second layer to employ the second-order pooling mechanism upon passing through the completely connected/linear layer.

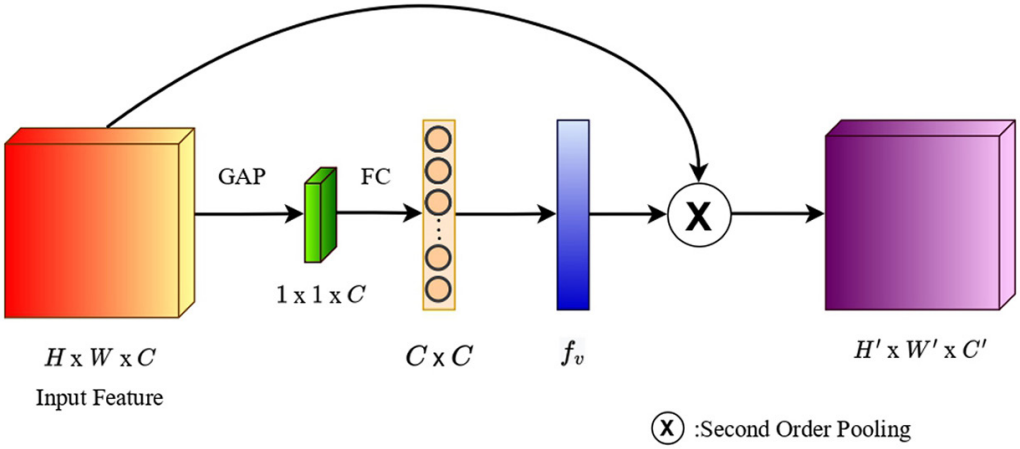

Fig. 2. The proposed module for streamlined attention blocks. The GAP and FC layers are the first to pass through the input feature. An SOP operation is carried out on the acquired feature vector (FV), with the instituted input feature generating the next block of feature mappings

As previously discussed, the contained spatial region of a particular feature map is how the feature refinement is achieved. As in [34], inter-spatial relationships are used to demonstrate this. By giving information that is easy to understand, the resulting spatial map has demonstrated its effectiveness.

Spatial attention is primarily concerned with "where" the visual attention should be drawn to extract finer features. The spatial structure has been presented as 2-dimensional R^{1*H*W} which is confined to the respective feature maps. Therefore, the spatial distribution of the refined features is obtained from the Linear/FC layer, where the aggregation of the input takes place, making sure to apply a second-order pooling operation. Here, the obtained spatial attention is observed due to the combination of the two relevant feature maps. In Section "Importance of second-order pooling", the significance of second-order pooling is discussed.

Importance of second-order pooling. The primary objective of convolutional neural networks is to capture numerous classes and objects specified in high-dimensional space. The model is extremely efficient whenever the higher-order representations are effectively and efficiently recognized by having the ability to enhance the non-linear modeling. Global second-order pooling (GSoP) is one such technique that has been shown to capture higher-order representations (GSoP) successfully.

Each dataset's sample pictures have been shown in Figure 3.

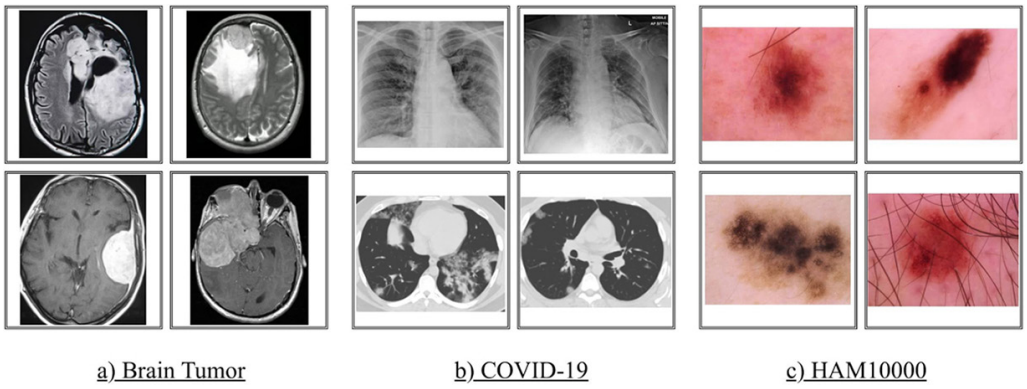

Fig. 3. Sample images of brain tumor, COVID-19, and HAM10000 datasets concerning both binary and multiclass (for the available*)

The success of GSoP layers in exhibiting realistic picture representations in the form of covariance matrices is shown in [38–40], resulting in state-of-the-art tasks like object identification, recognition, multimedia categorization, and fine cognitive recognition. In some of the older models, such as B-CNN and DeepO (2P), GSoP exhibits good performance and produces state-of-the-art results when the convolutional neural networks are trained from beginning to end, and the GSoP layer is included as the final layer or at the end of the entire model. To demonstrate the most recent findings, [41] was successful enough to consider the GSoP layer after each layer in the encoder component. By incorporating this mechanism into the design of the suggested simplified attention block module, we have demonstrated outcomes for the model. In Section "The proposed streamlined attention block module", the structure and interpretation of the suggested architecture are thoroughly demonstrated. The structure of the streamlined attention model embedded in Residual Neural Networks has been demonstrated in Figure 4.

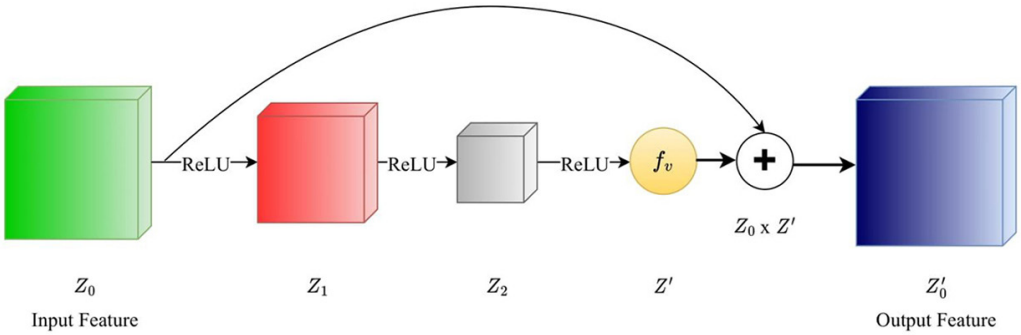

Fig. 4. Attention block module embedded into the ResNet architecture after each layer. The obtained feature vector fv undergoes a process of feature weight summation w.r.t Z0 and Z′, to obtain the succeeding output feature Z′0

The proposed streamlined attention block module. The bottleneck architecture of each third convolutional block in the corresponding encoder, ResNet, is coupled via an extended function in this methodology. We extend the typical architecture of the residual networks with a second, more functional attention block module that delivers careful neural attention for procuring long-range dependencies with less complexity than has previously been reported in the literature. To demonstrate the averaged weights for the subsequent layers of the pertinent attention block module, we use the stated global average pooling in the prior level of the function of the condensed attention block module while considering the proportions of the supplied input. The channel-based dimensions' top layer would be linear and fully connected through which the successive weights would travel. An element-wise multiplication or second-order pooling, including the proven weights of the prior layers, would come next. The significance of second-order pooling has been discussed in Section "Importance of second-order pooling", which finally aids in understanding that the attention obtained from the attention provided is considered extremely efficient.

Assume that the initial input is Z and the following feature block from the completely connected layer is Z_α so that we can express the upcoming pooling process as

$$Z_\alpha \otimes Z = Z' \tag{1}$$

In theabove equation, $\otimes$ is the relevant designation of second-order pooling. The resulting feature output is therefore considered to be Z'.

The encoder is linked to the simplified attention module's block. Given that the feature vector generated by the proposed module is Z' and the mainstream input is Z_0. In residual neural networks, the process requiring the summation of the corresponding skip connections can be explained as

$$Z' \otimes Z_0 \tag{2}$$

In the linear layer, the non-trained neuron is also newly trained during back-propagation, which is thought to be the reason for explicit attention. When data is supplied into the training phase, this non-linear mapping using the FC layer finally becomes trained when data is provided during the training phase, enabling it to comprehend the complex connections and, once more, map these traits to future convolution-layered features.

The neurons in the fully-connected layer have been trained from scratch during back-prop to gain explicit attention. When we feed the data, this trained layer gains a tendency to simultaneously interpret the complex relations as well as map the same to the layered convolutions. The Linear/FC layer helps in acquiring relevant features and projecting attention over the similarities. No specific requirement exists for a designated hyper-parameter to gain attention under this paradigm. It is considered ineffective or more complex to consider a specific hyper-parameter like ECANet. Hence, without applying external hyper-parameters, we have produced a comprehensive method. The weights updated during back-prop learn where to focus on the whole feature map due to the second-order pooling mechanism, without any loss of information, producing consistency concerning features. These insights helped the model outperform its competitors, and the results are discussed in the section below.

5 RESULTS AND DISCUSSION

Experiments employing the proposed attention model are carried out on a PC with a GPU (Graphical Processing Unit) NVIDIA RTX A4000 graphic card built on 16GB VRAM with 6144 CUDA cores. We focused on three key medical conditions for testing our method: COVID-19, brain tumors, and skin cancer. The model is trained and evaluated for the available datasets on both binary and multiclass datasets. This eventually aids in acquiring a collection of traits capturing the attention-partial. To be noted, we have used 5-fold cross-validation for our study as it is a gold standard for evaluation, and we denote the results in mean ± std format.

Our research's major goal is to shed light on hybrid convolutional networks that were approached from an attention-based standpoint, as there was less literature under tasks addressing medical-based data in this specific approach. Additionally, it is believed that a model should focus on how well it interprets the data rather than how accurate it is, with an accuracy range of 'x,' as addressed in [42]. This data interpretation method is only achievable by focusing on particular traits and giving them equal weights in the intermediate layers of the model to understand how better the attention module operates in our model. With this in mind, we can confidently state that the designed model offers reasonable visual attention.

Table 1. The table below provides accuracy acquired on individual datasets concerning COVID-19 (Binary), COVID-19 (Multiclass), Brain Tumor (Binary), Brain Tumor (Multiclass), and HAM10000. "N/A" indicates the unavailability of that specific kind of dataset in work and "–" indicates that the corresponding model wasn't developed to experiment on the mentioned data

Method	Datasets				
	COVID-19 (Binary)	COVID-19 (Multiclass)	Brain Tumor (Binary)	Brain Tumor (Multiclass)	HAM10000
Heidari et al. [41]	N/A	93.9%	–	–	–
Elaziz et al. [42]	96.09%	N/A	–	–	–
Panwar et al. [25]	94.04%	N/A	–	–	–
Khuzani et al. [12]	N/A	94%	–	–	–
Soares et al. [11]	97.38%	N/A	–	–	–
Banerjee et al. [26]	**98.93%**	N/A	–	–	–
Saxena et al. [30]	–	–	95%	N/A	–
Kang et al. [28]	–	–	90.35%	87.88%	–
Rehman et al. [27]	–	–	N/A	**98.69%**	–
Alanazi et al. [29]	–	–	**98.33%**	91.62%	–
Hsin-Wei et al. [20]	–	–	–	–	85.8%
FarhatAfza et al. [22]	–	–	–	–	93.4%
Khan et al. [21]	–	–	–	–	88.5%
ArdanAdi et al. [19]	–	–	–	–	78%
Our Model	98.06 ± 1.38	**94.75 ± 1.07**	73.88 ± 0.91	83.31 ± 4.71	**94.31 ± 0.91**

The GradCAM paper has profoundly impacted the field of deep learning by introducing a highly innovative visual explanation technique. This technique, referred

to as Gradient-weighted Class Activation Mapping (GradCAM), enables the user to interpret and evaluate the decisions of deep learning models through a series of heat maps that represent the contribution of each neuron to the overall prediction. This effectively allows for identifying the most important features of the model's predictions, enabling a greater understanding of its behavior and a more informed assessment of its performance. Thus, GradCAMs are used to demonstrate the suggested model's performance and attentive capacity to make its interpretability more visible. This helps prove the enhanced interpretations of the neural nets over applying a streamlined attention mechanism. Extensive experimentation was used to assess the results. Each value was tested for convergence using different settings. After repeatedly running the model for them, we consciously decided to obtain the mean and standard deviation for each execution. Every model underwent distinct amounts of training on each data set, with the training for each model being changed until convergence was attained. The images of each dataset were isotopically reshaped to 224×224. The image resolution concerning medical datasets is considerably higher compared to that of the generic dataset, so a batch size of 128 has been chosen to train the respective medical images conveniently.

While training, we initially implied to work on generic architectures such as ResNets, VGGNets, etc. But as addressed earlier in Section 3, there were multiple works about such experimentation. Hence, the scope of attention mechanism proved considerable results in multiple fields of study, as shown by our experimentation on generic data and fine-grained visualization-based data—CIFAR-10, CIFAR-100, and Aircraft's dataset, respectively. Therefore, the model addressed the medical data classification task from such a perspective.

As mentioned earlier, we the authors initially wanted to check on how the finer representations were attained on the model rather than preferring accuracy as our only metric. Therefore, the authors have tried to implement various types of medical data. This is a considerable reason for choosing COVID-19 (Binary), COVID-19 (Multiclass), Brain Tumor (Binary), Brain Tumor (Multiclass), and HAM10000. Most of the methods portrayed in Table. 1 corresponding to their model's accuracy are either ensemble learning-based methods or transfer learning-based methods. As it is already evident that these methods involve feature-based aggregation or model-based aggregation, they eventually stand out to be computationally expensive and less deployable, even if the presented accuracy stands out to be high.

The results obtained by this model (displayed in Table. 1) are considerably high, taking into account the performance portrayed by other models. This not only involves a rise in accuracy but is also considered to be less computationally expensive. Applying ensemble-based or transfer learning based on attention-based models would result in much better performance. This adds to our motive of model-based interpretation rather than focusing on raising the value of accuracy, as it is not only the evaluation metric to be concentrated on.

The confusion matrices of our model-based performance were obtained to interpret the rate of true-positives and true-nNegatives. The confusion matrices concerning each dataset have been displayed in Figure 5. The obtained results proved and gave enough scope for using attention-based mechanisms for proving results on medical imagery. We could sharply interpret the proposed model's range according to this method, which permitted us to emphasize the attention-partials in an initiated image.

During experimentation, we chose our model's optimizer to be Adam with a learning rate of 10^{-3} [43]. In the difficulty of predicting medical images, being computationally efficient and having a workable application on sparse data would both

be favorable. Adam is seen to converge more quickly and tends to reduce computing time. To make sure that the model doesn't over-fit, we have considered applying an early stopping mechanism. We have set early stopping to work for every 20 epochs when it observes a converging trend during training. The overall model's training was set to run for 75 epochs, whereas due to early stopping, most of the executions were terminated between a range of 50–60 epochs. The model enhances its ability to learn from epoch to epoch.

For effective backpropagation of the corresponding gradients in a multi-class classification problem, we have used categorical cross-entropy as our loss function i.e.

$$loss = \sum_{i=1}^{OP\ size} y_i.log.\hat{y}_i \tag{3}$$

Where $\hat{y}_i$ the ith is the scalar value of the respective model's output, and y_i is the associated target value. The output size, known as the "OP size," is determined by how many scalar values the model output contains. Class activation maps have demonstrated how the anticipated classes can be visualized (CAMs). These class activation maps assist in focusing on the particular object associated with the predicted class, allowing the user to discover the expected chunk without specifically retrieving it through a specific object. They aid in exhibiting the class that the particular network plans to perform. By aggregating the feature maps from the final layer, the CAMs show how well a model can draw attention to itself visually. Heat maps are another name for this collection of feature maps, as noted concerning [44–46].

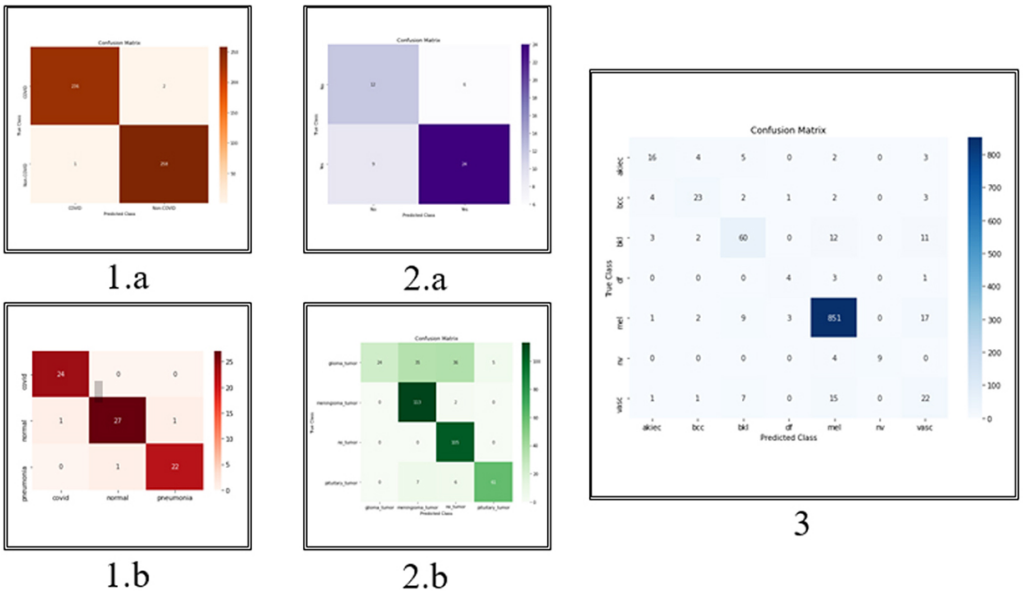

Fig. 5. Confusion Matrices obtained on COVID-19, Brain Tumor, and HAM10000 datasets where 1.a is COVID-19 (Binary), 1.b indicates COVID-19 (Multiclass), 2.a indicates Brain Tumor (Binary), 2.b indicates Brain Tumor (Multiclass) and 3 indicates HAM10000

Heatmaps show the concentrated region in red and help to visualize the image's partial localization. In plain language, the color variation ultimately indicates the area of the image that the model is paying the most attention to; for example, the redder the area, the more focused the model is on that area. The application of discriminative localization in the proposed system is particularly crucial in showcasing the potency of the acquired attention-based output because our system is based on

the attention mechanism, which causes the amount of concentration over a given region to be quite focused. As a result, the testing data is purposefully mixed, two images from each dataset are chosen at random, and the results are presented in Figure 6.

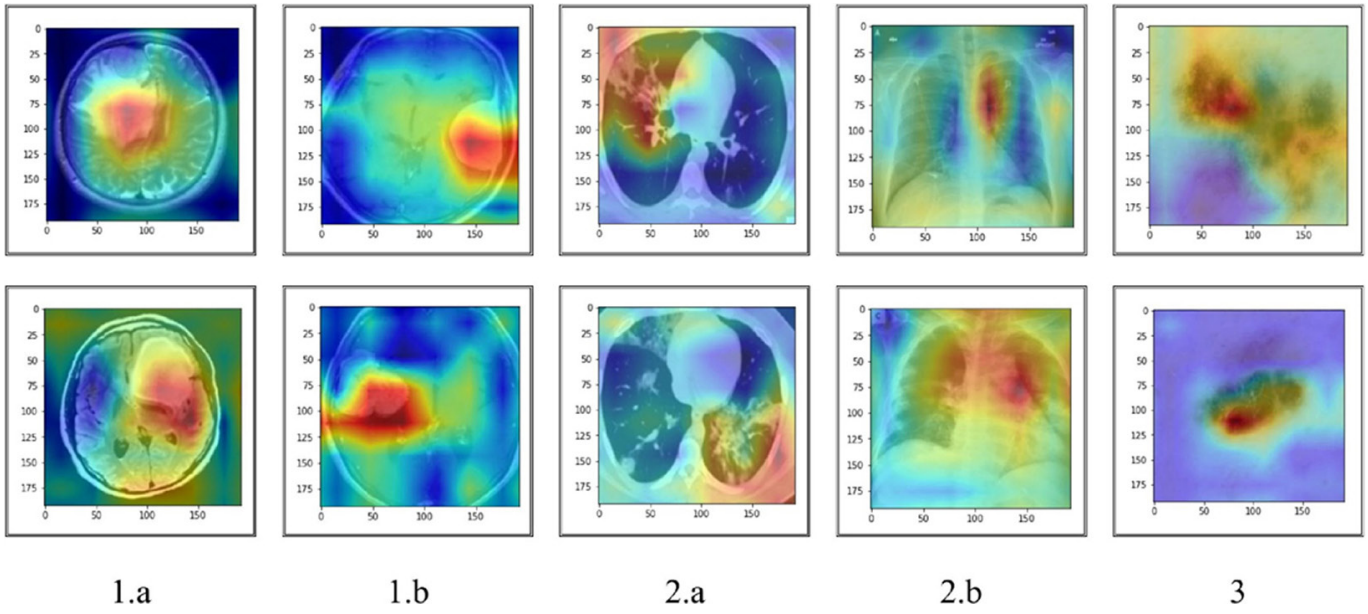

Fig. 6. The above images are the Class Activation Maps obtained for our proposed streamlined attention module on the respective medical data where 1.a is COVID-19 (Binary), 1.b indicates COVID-19 (Multiclass), 2.a indicates Brain Tumor (Binary), 2.b indicates Brain Tumor (Multiclass) and 3 indicates HAM10000

It can be interpreted that the model is acquiring the features from the various data mentioned and providing where it is proving attention. Thus it is visible that it can acquire the tumor proportions and provide attention to the tumor, and hence it can perform superiorly to the other methods.

According to the results, it can be seen that the model prefers to offer intense visual attention to the relevant medical facts and can perform better with extensive visual recognition. The future scope is to build a model that can deliver cutting-edge results for the explored datasets and other variations of medical data sets. The future scope of this work has been discussed in Section 6.

6 CONCLUSION AND FUTURE WORK

The findings show that the model outperforms extensive image perception and prefers to pay close attention to pertinent medical details. Using experiments, we observed that our model was able to surpass most of the standard works but was not able to generalize to brain tumor data sets. We initially developed this model with the question of how better hybrid convolutional neural networks would work with an attention block embedded in them. Obtaining considerable results on both generic and medical data motivates us to experiment with such a methodology on self-attention models.

This experimentation would let us interpret the patterns acquired by self-attention models in the absence of convolutions and also let us understand the difference between such models based on the results obtained via reason-based evaluation. Therefore, the scope of this work would extend to implementing such data on self-attention-based models i.e., on the ViTs – Vision Transformers. It is always better

that a model can learn the nuanced features of the data using self-attention rather than providing guided attention using several mechanisms. Thus, we look forward to designing a novel transformer network where the self-attention mechanism is implemented to acquire rich visual features without any external supervision.

7 ACKNOWLEDGMENT

The resources and computing environment was provided by the GITAM Institute of Technology. We are thankful for their support.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

8 REFERENCES

[1] Y. Y. Lecun and G. Bengio, "Deep learning," *Nature*, vol. 521, no. 2, pp. 436–444, 2015. https://doi.org/10.1038/nature14539

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017. https://doi.org/10.1145/3065386

[3] X. H. Kaiming, S. Zhang, and J. Ren, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[4] A. K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2015.

[5] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017. https://doi.org/10.1038/nature21056

[6] M. O. Gozes, H. Frid-Adar, D. Greenspan, and H. Browning, "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis," *Electrical Engineering and Systems Science*, vol. 1, no. 1, pp. 1–22, 2020.

[7] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized Medical Imaging and Graphics. Computer-aided Diagnosis (CAD) and Image-guided Decision Support*, vol. 31, pp. 198–211, 2007. https://doi.org/10.1016/j.compmedimag.2007.02.002

[8] H. K. Munir, A. Elahi, F. Ayub, and A. Frezza, "Cancer diagnosis using deep learning: A bibliographic review," *Cancers*, vol. 11, no. 9, pp. 1–36, 2019. https://doi.org/10.3390/cancers11091235

[9] M. A. Khan *et al.*, "COVID-19 case recognition from chest CT images by deep learning, entropy-controlled firefly optimization, and parallel feature fusion," *Sensors (Basel)*, vol. 21, no. 21, p. 7286, 2021. https://doi.org/10.3390/s21217286

[10] C. P. T. Schandl and H. Rosendahl, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018. https://doi.org/10.1038/sdata.2018.161

[11] P. E. Soares, S. Angelov, M. H. Biaso, and D. K. Froes, "Sars-cov-2 ct-scan dataset: A large dataset of real patientsct scans for sars-cov-2 identification," *MedRxiv,* vol. 3, pp. 8–16, 2020.

[12] Z. Khuzani, M. Heidari, and S. A. Shariati, "Covid-classifier: An automated machine learning model to assist in the diagnosis of covid-19 infection in chest x-ray images," *Scientific Reports,* vol. 11, no. 1, pp. 56–62, 2021. https://doi.org/10.1038/s41598-021-88807-2

[13] Kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/navoneel/brain-. [Accessed: 28-Aug- 2023].
[14] Kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/sartajbhuvaji/brain-. [Accessed: 28-Aug-2023].
[15] D. D. Himabindu and P. S. Kumar, "A streamlined attention mechanism for image classification and fine-grained visual recognition," *Mendel*, vol. 27, no. 2, pp. 59–67, 2021. https://doi.org/10.13164/mendel.2021.2.059
[16] K. D. Bahdanau and Y. Cho, "Neural machine translation by jointly learning to align and translate CoRR," in *Proc. The International Conference on Learning Representations (ICLR)*, San Diego, United States, 2016, pp. 1–15.
[17] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015. https://doi.org/10.18653/v1/D15-1166
[18] I. A. Nugroho, "Skins cancer identification system of haml0000 skin cancer dataset using convolutional neural network," *International Conference on Science and Applied Science (ICSAS)*, 2019, pp. 33–41. https://doi.org/10.1063/1.5141652
[19] B. H. W. Huang, C. H. Hsu, and S. Lee, "Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers," *The Journal of Dermatology*, vol. 48, no. 3, pp. 310–316, 2021. https://doi.org/10.1111/1346-8138.15683
[20] Y. Y. Ren, S. Long, J. Tian, and Z. Cheng, "Serial attention network for skin lesion segmentation," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 2, pp. 799–810, 2021. https://doi.org/10.1007/s12652-021-02933-3
[21] G. K. Young, B. Booth, R. Simpson, and S. Dutton, "Deep neural network or dermatologist?" in *Proc. iMIMIC/ML-CDS@MICCAI*, 2019. https://doi.org/10.1007/978-3-030-33850-3_6
[22] T. M. A. Khan, Y. D. Akram, and M. Zhang, "Attributes based skin lesion detection and recognition: A mask rcnn and transfer learning-based deep learning framework," *Pattern Recognition Letters*, vol. 143, pp. 58–66, 2021. https://doi.org/10.1016/j.patrec.2020.12.015
[23] M. F. Afza, M. A. Sharif, U. Khan, and H. S. Tariq, "Multiclass skin lesion classification using hybrid deep features selection and extreme learning machine," *Sensors*, vol. 22, no. 3, 2022. https://doi.org/10.3390/s22030799
[24] A. M. Rahimzadeh and S. M. Attar, "A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset," *Biomedical Signal Processing and Control*, vol. 68, pp. 21–41, 2021. https://doi.org/10.1016/j.bspc.2021.102588
[25] P. H. K. Panwar, M. K. Gupta, R. M. Siddiqui, and P. Menendez, "A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images," *Chaos, Solitons Fractals*, vol. 140, pp. 156–168, 2020. https://doi.org/10.1016/j.chaos.2020.110190
[26] R. Banerjee, V. Bhattacharya, P. K. Bhateja, and A. L. Singh, "Cofe-net: An ensemble strategy for computer-aided detection for covid-19," *Measurement*, vol. 187, no. 1, pp. 1–14, 2022. https://doi.org/10.1016/j.measurement.2021.110289
[27] A. Raza, H. U. R. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, "Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction," *PLoS ONE*, vol. 17, no. 11, p. e0276525, 2022. https://doi.org/10.1371/journal.pone.0276525
[28] S. Rehman, M. I. Naz, F. Razzak, and M. Akram, "A deep learning-based framework for automatic brain tumors classification using transfer learning," *Systems, and Signal Processing*, vol. 39, pp. 757–775, 2019. https://doi.org/10.1007/s00034-019-01246-3
[29] J. Kang, Z. Ullah, and J. Gwak, "MRI-based brain tumor classification using ensemble of deep features and machine learning classifiers," *Sensors (Basel)*, vol. 21, no. 6, p. 2222, 2021. https://doi.org/10.3390/s21062222

[30] M. M. F. U. Alanazi, S. J. Ali, A. Hussain, and M. Zafar, "Brain tumor/mass classification framework using magnetic-resonance-imaging-based isolated and developed transfer deep-learning model," *Sensors*, vol. 22, no. 1, 2022. https://doi.org/10.3390/s22010372

[31] A. P. Saxena, S. Maheshwari, and S. Tayal, "Predictive modeling of brain tumor: A deep learning approach," *ArXiv*, 1911.

[32] Z. Sohrabi, N. O. Alsafi, M. Neill, and A. Khan, "World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19)," *International Journal of Surgery*, vol. 76, no. 1, pp. 71–76, 2019. https://doi.org/10.1016/j.ijsu.2020.02.034

[33] Z. T. Ai, H. Yang, C. Hou, and C. Zhan, "Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (covid-19) in china: A report of 1014 cases," *Radiology*, vol. 296, no. 2, pp. 55–70, 2020. https://doi.org/10.1148/radiol.2020200642

[34] "Types of cancer," *Cancer.net*. [Online]. Available: https://www.cancer.net/cancer-. [Accessed: 28-Aug-2023].

[35] J. S. Woo, J. Park, and I. Lee, "Cbam: Convolutional block attention module," in *European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

[36] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," *ArXiv [cs.CV]*, 2019. https://doi.org/10.1109/CVPR42600.2020.01155

[37] L. J. Hu and G. Shen, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, 2018, pp. 7132–7141.

[38] J. P. Li, Q. Xie, and Z. Wang, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, 2018, pp. 947–955.

[39] F. Y. Cui, J. Zhou, X. Wang, and Y. Liu, "Kernel pooling for convolutional neural networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 3049–3058. https://doi.org/10.1109/CVPR.2017.325

[40] M. Y. Wang, J. Long, and S. Wang, "Spatiotemporal pyramid network for video action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2097–2106. https://doi.org/10.1109/CVPR.2017.226

[41] J. Z. Gao, Q. Xie, and P. Wang, "Global second-order pooling convolutional networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, 2019, pp. 3019–3028. https://doi.org/10.1109/CVPR.2019.00314

[42] M. Heidari, S. Mirniaharikandehei, A. Z. Khuzani, and G. Danala, "Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms," *International Journal of Medical Informatics*, vol. 144, pp. 104284–104284, 2020. https://doi.org/10.1016/j.ijmedinf.2020.104284

[43] E. A. Mohamed, K. M. Elaziz, A. Hosny, and M. Salah, "New machine learning method for image-based diagnosis of covid-19," *PLoS ONE*, vol. 15, pp. 1–18, 2020. https://doi.org/10.1371/journal.pone.0235187

[44] B. Wang, J. Zhao, R. Zhao, and C. Wang, "An adaptive method for topology optimization subjected to stationary stochastic forced excitations," *Eng. Optim.*, vol. 55, no. 2, pp. 177–196, 2023. https://doi.org/10.1080/0305215X.2021.1992613

[45] M. R. R. Selvaraju, A. Cogswell, R. Das, and D. Vedantam, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 618–626. https://doi.org/10.1109/ICCV.2017.74

[46] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018. https://doi.org/10.1109/WACV.2018.00097

9 AUTHORS

Dakshayani Himabindu Damineni received her B.Tech degree in Information Technology From JNTUH University, Hyderabad, India 2006 and M.Tech degree in Information Technology from IARE, JNTUH University, Hyderabad, India. Currently working as Assistant Professor, IT Department in VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India. Since 2014, Pursuing PhD from GITAM (Deemed to be University), Vizag, India. Her research interests are Deep learning, Machine Learning, Artificial Intelligence, Cloud Computing. She has lifetime membership in International Association of Engineers (IAENG) (E-mail: dakshayani.himabindu@gmail.com).

Dr. Praveen Kumar Sekharamantry is currently a post-doctoral researcher at University of Trento, Italy. Prior to his recent appointment at the Department of Information Engineering and Computer Science, Trento, Italy, he was a Faculty in Computer Science Engineering, GIT, GITAM University, Vizag, India for about fourteen years. Dr. Praveen received his PhD degree in 2019 from a reputed University as well as his Master's in Computer Science and Executive MBA degree from GITAM University. He acquired his Master's of Technology in Computer Science. Dr. Praveen Kumar Sekharamantry published a number of papers in preferred journals and chapters in books, and participated in a range of forums on deep learning, information security and Big Data. He also presented various academic as well as research-based papers at several national and international conferences. He is experienced in working on three international consultancy projects with exawizards Tokyo, Japan. He is a global certified professional on different technologies from Oracle and IBM. He is presently associated with a software company Bluetensor in Italy, working on projects of computer vision using machine learning (E-mail: psekharm@gitam.edu).

Dr. Rajakoti Badugu received his B.Tech degree in computer science & engineering from JNTUK University, Kakinada, India, in 2013 and M.Tech degree in information technology from GITAM University, Visakhapatnam, India, in 2015. He completed his Ph.D. degree in information technology from GITAM University, Visakhapatnam, India, in 2019. From 2019 till date, he is working as an Assistant Professor with the department of computer science & engineering, GITAM School of Technology, GITAM University, Visakhapatnam. His research interests include Cyber security, Computer networks security, Cryptography, Cloud computing and Blockchain technology. Dr. Raja also forms the part of Department's CARE committee where he liaisons with other Directorates and Schools in the University on various academic issues such as courses, Board of Studies and Session scheduling (E-mail: rbadugu@gitam.edu).