https://doi.org/10.31436/iiumej.v22i2.1752

CLASSIFICATION OF CHEST RADIOGRAPHS USING NOVEL ANOMALOUS SALIENCY MAP AND DEEP CONVOLUTIONAL NEURAL NETWORK

MOHD ADLI MD ALI[1*], MOHD RADHWAN ABIDIN [2], NIK ARSYAD NIK MUHAMAD AFFENDI[3], HAFIDZUL ABDULLAH[1], DAANIYAL REESHA ROSMAN[1], NU'MAN BADRUD'DIN[1], FAIZ KEMI[1] AND FARID HAYATI[1]

[1]Department. of Physics,
[2]Department of Radiology,
[3]Department of Internal Medicine,
International Islamic University Malaysia, Kuantan, Malaysia

[*]Corresponding author: qunox@iium.edu.my

(Received: 27th August 2020; Accepted: 30th January 2021; Published on-line: 4th July 2021)

ABSTRACT: The rapid advancement in pattern recognition via the deep learning method has made it possible to develop an autonomous medical image classification system. This system has proven robust and accurate in classifying most pathological features found in a medical image, such as airspace opacity, mass, and broken bone. Conventionally, this system takes routine medical images with minimum pre-processing as the model's input; in this research, we investigate if saliency maps can be an alternative model input. Recent research has shown that saliency maps' application increases deep learning model performance in image classification, object localization, and segmentation. However, conventional bottom-up saliency map algorithms regularly failed to localize salient or pathological anomalies in medical images. This failure is because most medical images are homogenous, lacking color, and contrast variant. Therefore, we also introduce the Xenafas algorithm in this paper. The algorithm creates a new kind of anomalous saliency map called the Intensity Probability Mapping and Weighted Intensity Probability Mapping. We tested the proposed saliency maps on five deep learning models based on common convolutional neural network architecture. The result of this experiment showed that using the proposed saliency map over regular radiograph chest images increases the sensitivity of most models in identifying images with air space opacities. Using the Grad-CAM algorithm, we showed how the proposed saliency map shifted the model attention to the relevant region in chest radiograph images. While in the qualitative study, it was found that the proposed saliency map regularly highlights anomalous features, including foreign objects and cardiomegaly. However, it is inconsistent in highlighting masses and nodules.

ABSTRAK: Perkembangan pesat sistem pengecaman corak menggunakan kaedah pembelajaran mendalam membolehkan penghasilan sistem klasifikasi gambar perubatan secara automatik. Sistem ini berupaya menilai secara tepat jika terdapat tanda-tanda patologi di dalam gambar perubatan seperti kelegapan ruang udara, jisim dan tulang patah. Kebiasaannya, sistem ini akan mengambil gambar perubatan dengan pra-pemprosesan minimum sebagai input. Kajian ini adalah tentang potensi peta salien dapat dijadikan sebagai model input alternatif. Ini kerana kajian terkini telah menunjukkan penggunaan peta salien dapat meningkatkan prestasi model pembelajaran mendalam dalam pengklasifikasian gambar, pengesanan objek, dan segmentasi gambar. Walau bagaimanapun, sistem konvensional algoritma peta salien jenis bawah-ke-atas kebiasaannya gagal mengesan salien atau anomali patologi dalam gambar-gambar

perubatan. Kegagalan ini disebabkan oleh sifat gambar perubatan yang homogen, kurang variasi warna dan kontras. Oleh itu, kajian ini memperkenalkan algoritma Xenafas yang menghasilkan dua jenis pemetaan saliensi anomali iaitu Pemetaan Kebarangkalian Keamatan dan Pemetaan Kebarangkalian Keamatan Pemberat. Kajian dibuat pada peta salien yang dicadangkan iaitu pada lima model pembelajaran mendalam berdasarkan seni bina rangkaian neural konvolusi yang sama. Dapatan kajian menunjukkan dengan menggunakan peta salien atas gambar-gambar radiografi dada tetap membantu kesensitifan kebanyakan model dalam mengidentifikasi gambar-gambar dengan kelegapan ruang udara. Dengan menggunakan algoritma Grad-CAM, peta salien yang dicadangkan ini mampu mengalih fokus model kepada kawasan yang relevan kepada gambar radiografi dada. Sementara itu, kajian kualitatif ini juga menunjukkan algoritma yang dicadangkan mampu memberi ciri anomali, termasuk objek asing dan kardiomegali. Walau bagaimanapun, ianya tidak konsisten dalam menjelaskan berat dan nodul.

KEYWORDS: *saliency mapping; chest radiograph; convolutional neural network*

1. INTRODUCTION

The convolutional neural network (CNN) has become the de-facto choice for image classification and object detection. It has shown that the network model can achieve human-level accuracy, including for medical images. Nevertheless, researchers are still finding ways to improve the classification performance with novel ideas. The majority of this research focuses on developing ever more complex and deep architecture. In this paper, we test the idea of changing the input typing rather than the model architecture. Instead of using a regular medical image, the saliency map is proposed to be the alternative input.

1.1 Introduction to Saliency Map

Itti et al. [1] introduced the concept of the saliency map in 1998. A saliency map is a numerical map that localizes an object (or objects) in an image that is deemed interesting (salient). In other words, the map emphasizes relevant features in an image while at the same time suppressing irrelevant features. Saliency maps have been employed in many tasks, including image classification, object detection, and image segmentation [2,3].

Methods for creating a salient map can be divided into the top-down and bottom-up approaches [4]. In the bottom-up approaches, the salient map is constructed based solely on the image's feature. Features such as color mapping, contrast, edges, and objection placement are used to localize the image's salient region. Famous bottom-up algorithms are Binary Normed Gradient for Objectness [5], the Fine-Grained [6], and Spectral Residual [7]. However, [8] stated that medical images produced by conventional modalities such as CXR, computer tomography (CT) scan, and ultrasound are mostly homogenous and possess very few color variants. In situations like this, most conventional bottom-up algorithms will fail to localize any salient object in the image; this is shown in Fig. 4.

Contradicting the previous method, the top-down approach produces a salience map based on the task given. The algorithm takes external cues from a human or model feedback to construct the final salience map. This method is fast becoming the mainstream solution, especially for medical images, as it can produce precise salient region boundaries even in the presence of shades or reflections [9]. However, since the techniques are based on a supervised CNN model, from which it naturally inherits CNN dependencies. First, it requires a large number of annotated samples for training purposes. Secondly, its development and deployment require access to accelerated hardware. These two requirements are an obstacle for the practical deployment of such technology in the medical

field, especially in Malaysia. Currently, Malaysia lacks any open medical image dataset, and very few hospitals are equipped with, or have access to, accelerated hardware.

1.2 Anomalous Saliency Mapping

In this paper, we introduce a new algorithm called the Xenafas algorithm that produces two novel anomalous saliency maps call Intensity Probability Map (IPM) and Weighted Intensity Probability Map (WPM). Different from the bottom-approach which only takes internal image cues, our approach takes cues from the probability mapping of a pixel's intensity relative to a cluster of similar images. However, it also does not require annotated samples or accelerated hardware to create the saliency map, which is practical in the context of Malaysia's clinical settings. Therefore, the algorithm can be considered as a middle ground between the bottom-up and top-down approach. We test the algorithm on a chest radiograph (CXR) dataset to see if the algorithm can create a salience region by highlighting pathological features such as air space opacities, masses, and foreign objects.

2. LITERATURE REVIEW

For readers who want more information on the saliency map, ref [8] provides an extensive review of the subject matter. This paper's literature review will focus on the application of the saliency map in medical image analysis.

The application of saliency in medical images can be separated into two categories, depending on when it is used. The majority of research only applied it post-training and solely for model interpretations; it is not actively involved in model training. For example, in [10], the saliency map produced by the class activation mappings (CAM) [11] is used to validate the feature selected by the CheXNeXt model for its classifications. Similarly, in [12], a saliency map is created via the guided back-propagation method [13], which is then used to provide interpretability for model classification on breast cancer image classifications. Research done by [14-16] also shows similar traits. However, it is vital to mention the finding by [17], in which the author demonstrated that most algorithms used to create this saliency map are inconsistent when repeated. Among all algorithms, the Grad-CAM [18] algorithm shows the most consistency. Thus, the trustworthiness of using a saliency map to validate clinical CNN models is questionable.

The second type of saliency map research actively uses it in the model training. For example, in [19], the saliency map in the form of an attention map reduces a model false-positive rate. Similar to a saliency map, the attention map produced by the Attention Gate (AG) algorithm suppresses irrelevant regions in the image. In [20], localization of pulmonary lesions in CXR images is achieved by extracting a saliency map from a CNN model. Likewise, in [21], a saliency map is generated and used to detect polyps in capsule endoscopy. Various bottom-up saliency algorithms are used for segmenting skin cancer in [22,23].

To the best of the authors' knowledge, there is yet a paper that examines the effect of using a saliency map as input for chest radiograph classification. Therefore, in this paper, we tested the effect of using the proposed saliency map, IPM, and WPM, which will enhance the classification performance of the CNN model. In addition, we also test if the proposed saliency map successfully highlights all pathological features in a CXR image. For the interested reader, a review on the classification of CXR by supervised CNN models can be found from [24].

https://doi.org/10.31436/iiumej.v22i2.1752

3. METHODOLOGY

3.1 The Xenafas Algorithm

We proposed the Xenafas method, an algorithm that indicates anomaly regions' location on a CXR image, based on the likelihood of a pixel's intensity (opacities) at a given location. The method starts with creating a control dataset. Images for this dataset must be cherry-picked; avoiding images containing any form of anomalies. Examples of anomalies include but are not limited to; any pathology, foreign body, extreme variation such as dextrocardia, rotated film, and patients in non-standard body positions.

After the control dataset has been created, the images are clustered into several groups using the K-Means algorithm. This step is needed to address the variation in patient body shape, image quality between x-ray machines, and the patient's body's orientation when the x-ray image is taken. The number of clusters, K, depends on the homogeneity of the images in the dataset. A good homogenous dataset will use a K value of 1–3, while a heterogeneous dataset will use a value between 7–10. Next, the 2D pixel intensity distribution or ProbMat is created as shown in Algorithm 1.

```
Data: Batch of images
Result: 2D Pixel Intensity Distribution (probMat)
initialization;
probMat = [ ][ ];
for x in image width, y in image height, do
    pxiLs = [ pixel intensitity at (x,y) for image in image batch ];
    probFunc( ) = non-parametric probability function of pxiLs;
    probMat[ x ][ y ] = [ probFunc( i ) for i in range 0 to 255 ];
end
```

Fig. 1: The pseudocode for producing the ProbMat.

There are several ways to create a non-parametric probability function; one of the most popular is to use the Kernel Density Estimation method (KDE). KDE is easy to implement; however, computationally intensive when it is scaled to sample high-resolution images. A dataset with images with 256 by 256 resolution will need 65 536 KDE modeling to create all the necessary ProbMat. This requirement will quickly exhaust the memory resource of a computer. Additionally, there is no clear guideline on determining the appropriate bandwidth value of a KDE model.

We proposed another method as an alternative to KDE. In this method, we first create a histogram of pixel intensity for all CXR images in the subcluster K, at a specific value of x and y. From the histogram, a discrete probability function can be obtained. A continuous function for all possible intensity values is approximate by combining the discrete probability function with cubic spline interpolation. The Savitzky-Golay filter was then applied to smooth the probability distribution function further. Using this continuous probability function, it is now possible to create a matrix (ProbMat) representing the probability of all intensity values at any given location. Pseudocode shown Fig. 2 is used to create the anomalous saliency map. In this part, the ProbMat is used to produce the Intensity

Probability Map (IPM) and Weighted Intensity Probability Map (WPM) for all CXR images.

```
Data: Test image, ProbMat
Result: IMP or WMP Heatmap
initialization;
heatMap = [ ][ ];
for x in image width, y in image height do
    heatMap[ x ][ y ] = ProbMat[ x ][ y ][ test image pixel intensity at
    (x)(y)];
end
if output type == WMP then
    return wmpFunc( heatMap );
else
    return heatMap as IMP image;
end
```

Fig. 2: Pseudocode to produce IPM and WPM saliency map.

The WPM function is given by Eq. (1),

$$W(x,y) = P(i_{x,y})i_{x,y} \tag{1}$$

The weighted pixel intensity, W at position x, y is equal to the product of its intensity, i, and the intensity likelihood, $P()$, at the same locations. In WPM, the original pixel intensity acts as a weight for the likelihood. Thus, only anomaly regions with high opacities will be shown in WPM; lucent anomalies will be suppressed. In visualizing IPM and WPM images, pixels with lower likelihood will have a higher intensity (appear brighter) than pixels with high likelihood. Thus, a region that is marked brightly (highlighted) is a region that the algorithm considers to have anomaly features.

One of the IPM and WPM images' fundamental weaknesses is that it suppresses anatomical landmarks. Without anatomical landmarks, it is difficult to determine the location of an anomalous region relative to an organ. To solve this issue, we added IPM/WPM heatmap as a layer on top of the corresponding images. Though, only regions that exceed the Otsu threshold [25] are incorporated into the images. An example of IPM, WPM, Infused-IPM (IIPM) and Infused-WPM (IWPM) is shown in Fig. 5. For comparison purpose, Fig. 4 shows the output of conventional bottom-up saliency mapping algorithms called Fine-Grained and Spectral Residual. The implementation of both these algorithms is taken from the OpenCV.

3.2 Classification Method

This paper aims to test whether or not replacing CXR images with IPM and WPM will improve CNN's classification performance. Thus, to ensure any performance changes are due to the input type and not the CNN architectures, only familiar deep CNN models are used. Figure 3 shows the network architecture used, with the base model being MobileNet, DenseNet121, ResNet50, VGG19 and Xception [26-30]. The implementation of base model network architecture is taken from TensorFlow (Ver. 2) library and with the pre-trained weight from ImageNet [31].

https://doi.org/10.31436/iiumej.v22i2.1752

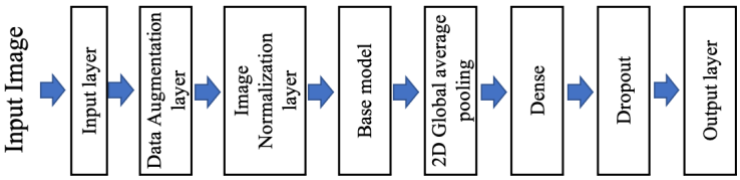

Fig. 3: Classification network architecture.

All models are trained with 100 epochs; however, an early stop is executed if there is no improvement in the loss value after ten epochs. All models were trained and tested on the Google Cloud platforms.

We have chosen the regular classification metric performance for model validation: precision, sensitivity, receiver operating characteristic curve (ROC-AUC), and the area under the precision versus sensitivity curve (PR-AUC).

3.3 Dataset

The dataset that is being used in this research is the Google-NIH dataset [32]. Initially, NIH provides the images while the labels are provided by Google [33]. It is important to note that only labels for the test and validation dataset are provided from the source. To create the training dataset for this study, we split the original validation dataset into a new training and validation dataset with a ratio of 0.3. All datasets are imbalanced datasets.

3.4 Qualitative Validations

To test the clinical relevance of the proposed algorithm, several normal and anomalous CXR images were selected, and the resulting IPM and WPM were examined qualitatively by a certified radiologist. Anomalous CXR that was chosen includes images of a rotated film, foreign body, cardiomegaly, and masses. For model interpretation, the Grad-CAM [18] algorithm is used to visualize which region of the CXR is relevant to the model when making the class classification.

4. RESULTS AND DISCUSSION

4.1 Classification Result

Table 1 shows several classification metrics obtained by various models and input-data types to classify the test dataset for air space opacity. While Table 2 shows similar metrics for the classification of CXR images with masses/nodules. Entries with the highest score for a particular metric is bolded.

The result obtained is not particularly easy to decipher. The highest score in PR-AUC, accuracy, and precision is obtained by ResNet50+Image, ResNet50+IIPM, and ResNet50+IWMP, respectively. Xception+Image and VGG19+Image do have a higher precision score, however, both results were rejected due to their sensitivity score being less than 0.5. This means the models falsely label the majority of positive samples. The model with the highest sensitivity score is VGG199+WPM, with a score of 0.930. However, the model precision is quite low, only 0.672, thus the next model, DenseNet121+IWPM, will be a better choice, having obtained 0.893 in sensitivity and 0.775 in precision. Meanwhile, DenseNet121+Image obtained the highest ROC-AUC score, 0.877.

https://doi.org/10.31436/iiumej.v22i2.1752

Table 1: Classification performance of CNN models with different base models and input data type for the airspace opacity dataset. ROC, PR, ACC, Pre, Sen stand for ROC-AUC, Precision-Recall area under the curve, accuracy, precision, and sensitivity respectively. %Δ represents the percentage difference in score compared to the model with image as input

Input Type	ROC-AUC	%Δ	PR-AUC	%Δ	Accuracy	%Δ	Precision	%Δ	Sensitivity	%Δ
MobileNetV2										
Image	0.844		0.883		0.770		0.814		0.781	
IPM	0.824	-2.4	0.865	-2.1	0.754	-2.1	0.787	-3.3	0.789	1.0
WPM	0.801	-5.2	0.835	-5.4	0.670	-13.0	0.851	4.4	0.522	-33.2
IIPM	0.846	0.2	0.879	-0.5	0.767	-0.5	0.821	0.8	0.763	-2.3
IWPM	0.841	-0.3	0.873	-1.2	0.772	0.3	0.785	-3.6	0.834	6.9
DenseNet121										
Image	**0.877**		0.903		0.776		0.878		0.712	
IPM	0.836	-4.7	0.874	-3.3	0.711	-8.4	0.878	-0.1	0.581	-18.3
WPM	0.820	-6.5	0.860	-4.8	0.736	-5.2	0.795	-9.5	0.733	3.0
IIPM	0.873	-0.4	0.905	0.2	0.763	-1.7	0.877	-0.1	0.686	-3.6
IWPM	0.874	-0.3	0.904	0.1	0.788	1.5	0.775	-11.8	0.893	25.4
Xception										
Image	0.856		0.890		0.683		0.921		0.495	
IPM	0.809	-5.6	0.851	-4.4	0.736	7.8	0.757	-17.9	0.803	62.1
WPM	0.797	-6.9	0.836	-6.1	0.729	6.7	0.797	-13.5	0.714	44.1
IIPM	0.835	-2.5	0.871	-2.1	0.732	7.2	0.856	-7.1	0.646	30.4
IWPM	0.841	-1.8	0.874	-1.8	0.723	5.7	0.876	-4.9	0.606	22.4
ResNet50										
Image	0.873		**0.907**		0.784		0.881		0.725	
IPM	0.838	-4.0	0.879	-3.1	0.757	-3.5	0.831	-5.7	0.728	0.4
WPM	0.829	-5.0	0.874	-3.7	0.759	-3.2	0.802	-9.0	0.775	6.9
IIPM	0.870	-0.4	0.906	-0.2	**0.790**	0.6	0.872	-1.0	0.745	2.8
IWPM	0.867	-0.6	0.905	-0.3	0.743	-5.3	**0.904**	2.6	0.621	-14.3
VGG19										
Image	0.840		0.876		0.654		0.910		0.447	
IPM	0.805	-4.2	0.860	-1.8	0.761	16.4	0.769	-15.5	0.841	88.2
WPM	0.796	-5.3	0.843	-3.8	0.697	6.5	0.672	-26.2	**0.930**	108.3
IIPM	0.858	2.2	0.894	2.1	0.760	16.1	0.875	-3.9	0.683	52.9
IWPM	0.845	0.6	0.887	1.3	0.760	16.1	0.852	-6.3	0.707	58.4

Next, we analyze if using the proposed anomalous saliency mapping as input will result in a better classifier for the airspace opacity dataset. We are particularly interested if such change in input can boost the performance of shallower CNN models (MobileNetV2 and DenseNet121) to comparable performance of deeper CNN models (ResNet50, VGG19 and Xception). What is evident from the result, using the alternative data types as input enhances the model's sensitivity. For example, VGG19+WPM, which obtained the highest sensitivity, obtained a 108.3% improvement compared to VGG19+IMG. This sensitivity improvement is more apparent in deep CNN models (ResNet50, VGG19, and Xception) than shallower CNN models (MobileNetV2 and DenseNet121).

As one might expect, any improvement in sensitivity tends to reduce model precision. Nevertheless, in most results, the degree of precision reduction is less than the degree of sensitivity gain. For example, the model Xception+IIPM obtained an increase of 30.4% in sensitivity while only reducing its precision by 7.1% compared to Xception+Image.

The answer to which CNN model and input data type perform the best, depends on the purpose of the model. For screening purposes, then DenseNet121+IWPM will be the recommended model as it obtained the second-best sensitivity score while maintaining a reliable precision score. For precise clinical classification, then ResNet50+IWMP is recommended.

Table 2: Classification performance of CNN models with different base models and input data type for the mass/nodule dataset. ROC, PR, ACC, Pre, Sen stand for ROC-AUC, Precision-Recall area under the curve, accuracy, precision, and sensitivity respectively. %Δ represents the percentage difference in score compared to the model with image as input

Input Type	ROC-AUC	%Δ	PR-AUC	%Δ	Accuracy	%Δ	Precision	%Δ	Sensitivity	%Δ
MobileNetV2										
Image	0.588		0.194		0.709		0.209		0.336	
IPM	0.568	-3.4	0.188	-3.2	0.729	2.8	0.201	-4.2	0.268	-20.2
WPM	0.588	0.0	0.204	5.2	0.755	6.5	0.206	-1.4	0.220	-34.3
IIPM	0.577	-1.7	0.182	-6.4	0.805	13.4	0.194	-7.1	0.095	-71.7
IWPM	0.573	-2.4	0.192	-1.0	0.659	-7.2	0.188	-10.2	0.383	14.1
DenseNet121										
Image	0.606		0.199		0.739		0.228		0.308	
IPM	0.587	-3.1	0.198	-0.6	0.716	-3.1	0.205	-10.1	0.308	0.0
WPM	0.607	0.3	0.209	4.9	0.541	-26.8	0.197	-13.8	**0.664**	115.4
IIPM	0.619	2.2	0.215	7.9	0.669	-9.4	0.222	-2.8	0.478	54.9
IWPM	0.617	1.9	0.216	8.2	0.731	-1.0	0.241	5.7	0.366	18.7
Xception										
Image	0.637		0.231		0.702		0.227		0.410	
IPM	0.587	-7.8	0.197	-14.7	0.633	-9.8	0.196	-13.8	0.464	13.2
WPM	0.575	-9.8	0.200	-13.5	0.611	-13.0	0.177	-22.0	0.437	6.6
IIPM	0.602	-5.5	0.206	-10.6	0.737	5.0	0.213	-6.4	0.278	-32.2
IWPM	0.609	-4.5	0.210	-9.2	0.570	-18.7	0.192	-15.5	0.580	41.3
ResNet50										
Image	0.643		0.242		0.689		0.225		0.437	
IPM	**0.644**	0.2	0.268	11.0	0.631	-8.4	0.222	-1.3	0.580	32.6
WPM	0.619	-3.7	0.253	4.5	0.556	-19.3	0.199	-11.4	0.647	48.1
IIPM	0.614	-4.5	0.210	-13.0	0.745	8.1	0.228	1.2	0.292	-33.3
IWPM	0.612	-4.7	0.222	-7.9	**0.820**	19.0	**0.299**	33.2	0.149	-65.9
VGG19										
Image	0.619		0.257		0.741		0.230		0.308	
IPM	0.623	0.6	**0.279**	8.5	0.545	-26.4	0.195	-15.3	0.647	109.9
WPM	0.561	-9.3	0.194	-24.6	0.733	-1.1	0.206	-10.7	0.271	-12.1
IIPM	0.620	0.1	0.245	-4.8	0.743	0.2	0.244	5.9	0.339	9.9
IWPM	0.628	1.4	0.260	1.0	0.752	1.5	0.248	7.7	0.319	3.3

It is worth noting that, since this dataset is imbalanced, the PR-AUC score is more important than the ROC-AUC score. DenseNet121+IWPM also obtains a PR-AUC score of 0.904, a

mere 0.003 less than the highest score, which is 0.907 by ResNet50+Image. In addition to obtaining a reliable classification score, it also has the advantage of requiring fewer computing resources than ResNet50, VGG19, and Xception. Thus, it is more practical to be deployed in Malaysian hospitals.

In Table 2, results show that all models failed to achieve acceptable classification performance for the mass/nodule testing-dataset. No model obtained a precision score of more than 0.5, meaning the majority of positive classifications were actually false. It is worth pointing out that all sensitivity scores for IMG input were lower than 0.5, thus all model missed the majority of the mass/nodule samples. Only DenseNet121+WPM, ResNet50+ WPM/IPM and VGG19+IPM manage to achieve a sensitivity score above 0.5.

4.2 Qualitative Assessment

Figure 4 shows the example of a saliency map produced by the Fine Grained and Spectral Residual algorithm, a conventional bottom-up algorithm. As shown in the figure, the Fine Grained failed to emphasize or suppress any feature in the image. Conversely, the Spectral Residual suppressed almost all featured, making it impossible to extract any meaningful information from its saliency map. Aligned with what was mentioned in [9], the conventional bottom-up saliency map algorithm cannot produce meaningful mapping for CXR images.

Fig. 4: The saliency map of a CXR image (a) produced by the Fine Grained, (b) and Spectral Residual (c) algorithms.

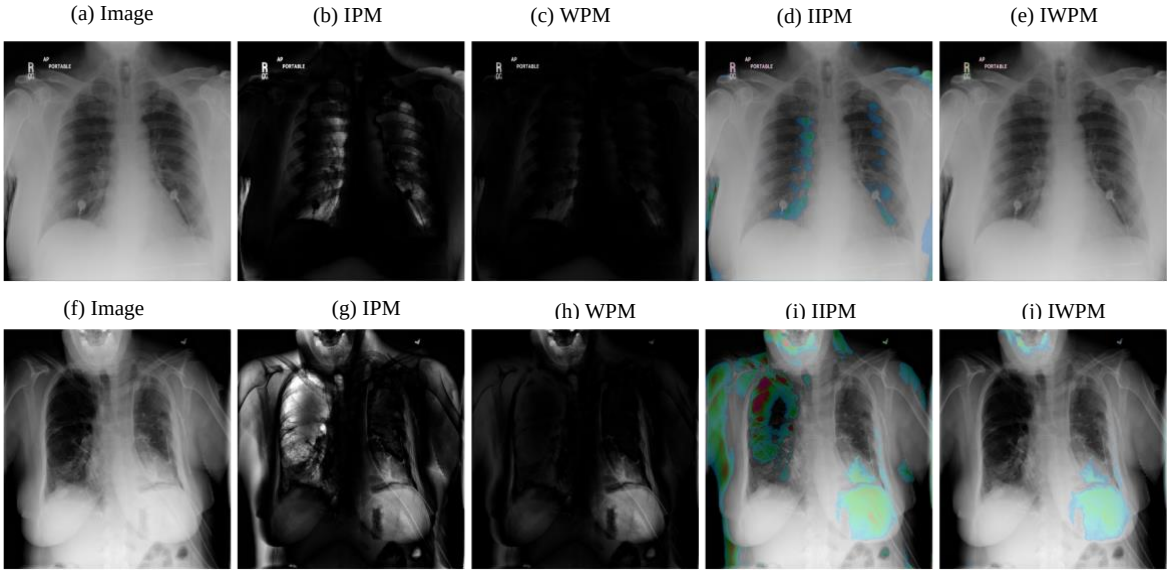

Fig. 5: A comparison between the original image and the corresponding IPM, WPM, IIPM and IWPM for a patient in a normal (a-e) and rotated (f-j) position.

https://doi.org/10.31436/iiumej.v22i2.1752

Figure 5(a) shows CXR images with no visible abnormalities; the resulting IPM and WPM are shown the Fig. 5(b) and 5(c), respectively. There is no apparent region highlighted for the WPM image, implying that the algorithm does not identify any anomaly in the original CXR image. However, the perihilar region is incorrectly highlighted in the IPM image; this may suggest that the IPM may be over-sensitive in highlighting anomalies in CXR images.

Next, we examine how the algorithm processed CXR images taken for an incorrectly positioned patient, or for a rotated x-ray film. For example, in Fig. 5(d), the patient's trachea is not located in the midline, suggesting that the patient may be rotated relative to the film. This orientation gives the appearance that the right lung is more lucent than the left. In the produced IPM image, the lucent region is highlighted, whereas WPM does not highlight this feature as WPM suppresses lucent anomalies. Whether or not the cause of the right lung lucency is significant is still an anomaly from the imaging perspective. Thus, the algorithm should highlight this anomaly as in the IPM image and then proceed to validate it by a radiologist.

Another feature that indicates that the patient is in an abnormal position is the presence of teeth in the CXR image. The anomaly is highlighted in the IPM and more evidently in the WPM image. Teeth usually are not presented in a CXR, thus it is a form of anomaly that should be highlighted by a UAS algorithm. However, the algorithm incorrectly highlighted the patient breasts; this error may have been caused by the lack of images containing breasts in the control dataset.

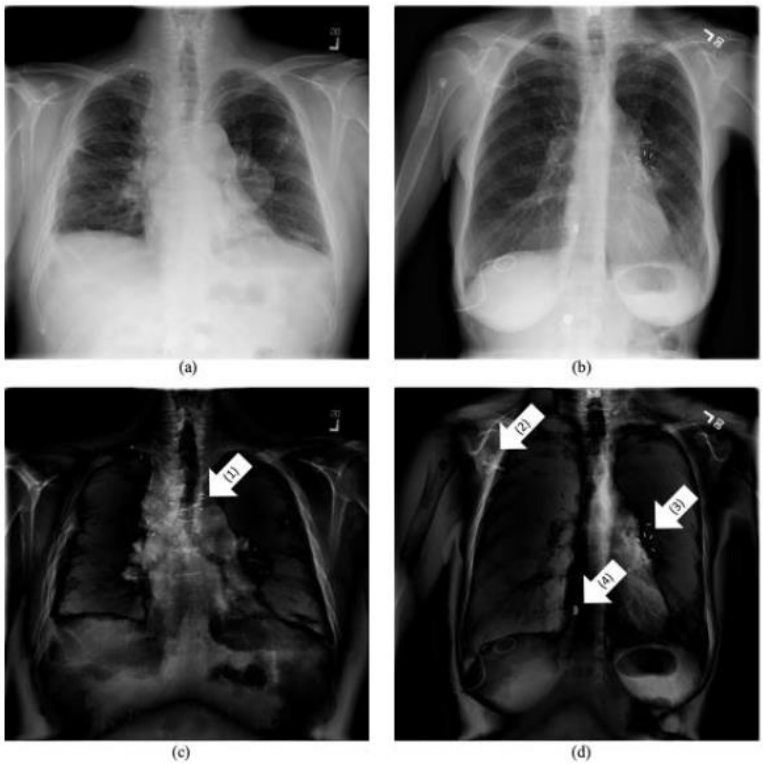

Fig. 6: Chest radiograph with foreign body, (a) and (b).
The resulting WPM images (c) and (d) clearly show the foreign bodies, arrow (1)-(4).

Figure 6(a) and 6(b) show CXR with foreign bodies, and the resulting WPM images are shown in Fig. 6(c) and (d). In both WPM images, the foreign bodies (arrows) are highlighted clearly and are more apparent than in the original CXR images. The opacity of biological

matter around the foreign body is suppressed, making it appear lucent. With this result, it can be assumed that WPM images may help in identifying foreign objects in CXR.

Next, the algorithm capabilities in highlighting pathological changes are demonstrated. Figure 7(a) shows a CXR with cardiomegaly, and it is highlighted clearly in both IPM, Fig. 7(b) and WPM Fig. 7(c) image. On the other hand, Fig. 7(d) shows CXR with a homogenous opacity at the right lower lung zone that does not obscure the cardiac border. This feature is not highlighted in both IPM, Fig. 7(e) and WPM, Fig. 7(f) images.

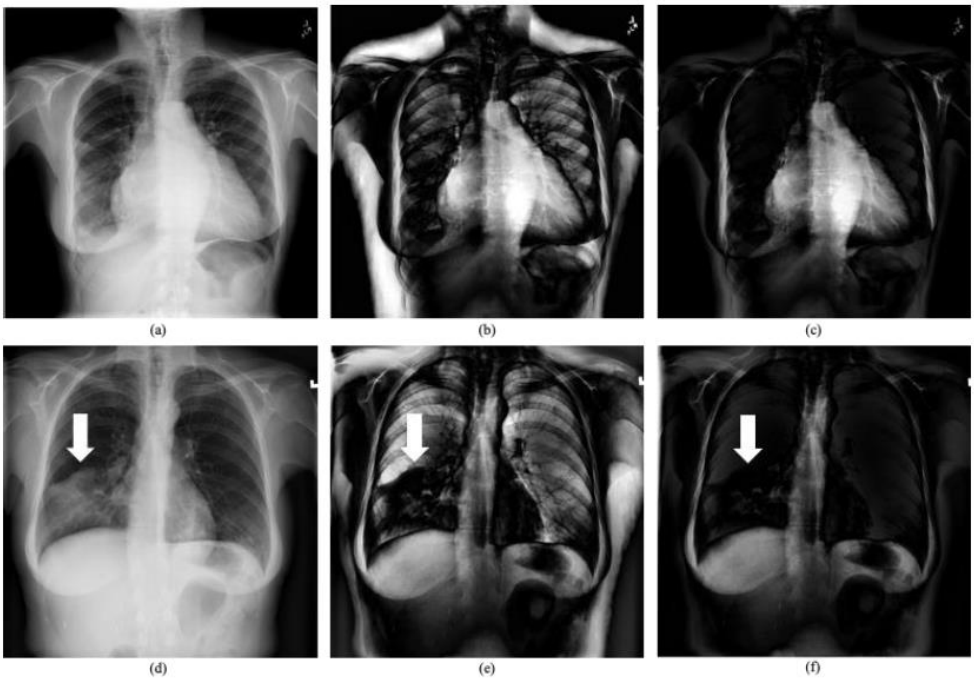

Fig. 7: The cardiomegaly feature in (a) is clearly highlighted in the WPM (b) and IPM (c) image. However the generalized opacity feature in (d) is absent in the IPM (e) and WPM (f) images.

Additionally, the algorithm also frequently failed to highlight opacity due to mass and nodules. The single nodule in Fig. 8(a) was not highlighted in the resulting WPM image, Fig. 8(d). The same can also be said for the multiple nodule-like opacities at bilateral mid and lower lung zones, as shown in Fig. 8(b). Only some of the lung masses are highlighted in resulting WPM images, Fig. 8(e). An example of a correctly highlighted lung mass is shown in Fig. 8(c) and 8(f).

4.3 Grad-CAM Results

To meaningfully deploy a developed model in clinical use, it must show some degree of interpretability and the classification must be validated based on some biological markers. For this reason, we use the Grad-CAM, [18], to visualize which region on the input data is emphasized. Figure 9(a) shows an example of CXR having airspace opacifications. Figure 9(b)-(f) shows the output of the Grad-CAM algorithm for DenseNet121 models that were trained with different input data types. The only models that were trained using the WPM and IIPM as input were correctly labeled the sample.

https://doi.org/10.31436/iiumej.v22i2.1752

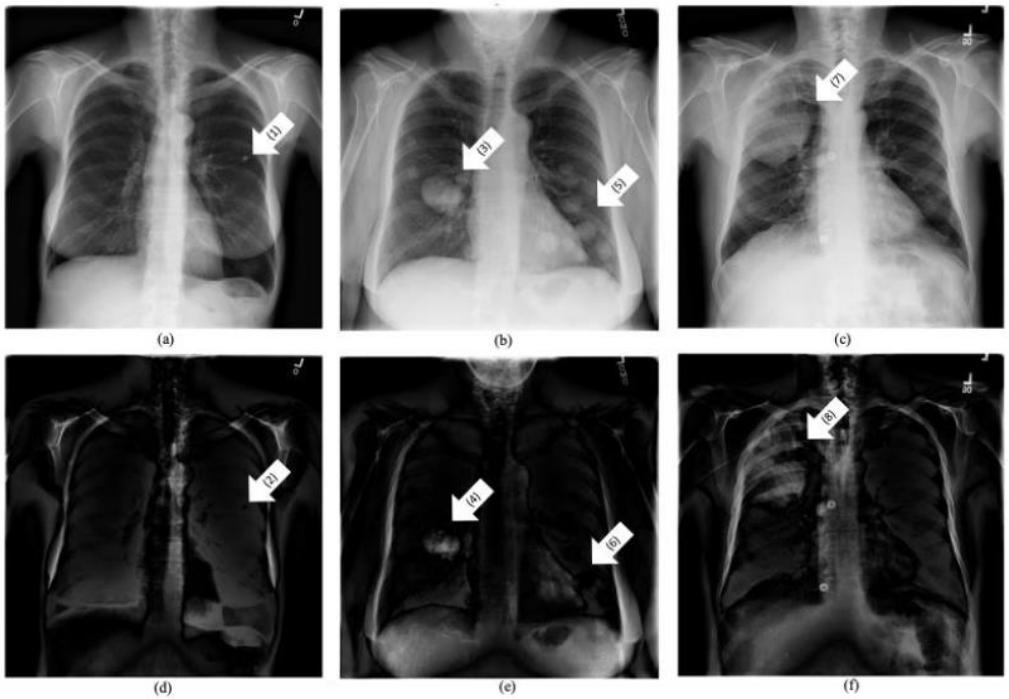

Fig. 8: Most of the lung nodules and masses in (a) and (b) failed to be highlighted in the resulting WPM images, (d) and (e). Only lung masses in (c) were successfully highlighted in (f).

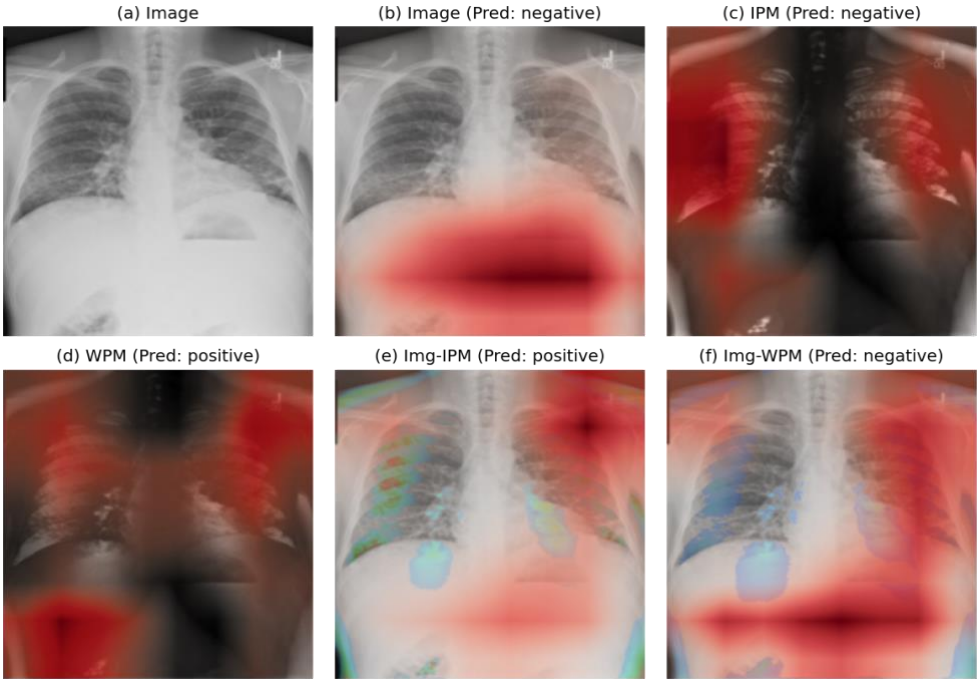

Fig. 9: Output of the Grad-CAM algorithm for air space opacification dataset. Regions that are marked red are regions that the DenseNet-121 deems important.

An obvious pattern that emerges in Fig. 9 is that models that received the original CXR image (image, IIPM, and IWPM) as input tend to mark the lower-left diagram. On the other hand, the model that takes IPM and WPM tends to focus more on the lung and shoulder region. It is not exactly certain why the DenseNet121-IPM model incorrectly labeled the

image even though it correctly emphasized the lung region. One reason that can be attributed is the model emphasized the right lung more than the left lung. Both models that correctly labeled the image, DenseNet121-WPM, and DenseNet121-IIPM, emphasize the left lung region. Even though DenseNet121-WPM also marked the left lung region; it emphasized more on the diagram, hence the wrong labeling.

Figure 10(a) shows a sample of CXR that is positive for mass. The location of the mass is in the left upper and lower lobe. Figure 10(b)-(f) show the output result of the Grad-CAM algorithm for different ResNet50 models that were trained by different input data types. Models trained using images, IIPM and IWPM, show similar marked regions; they extend from the left clavicle bone to the lung right middle lobe. For IIPM and IWPM, the region does not cover any mass. Thus, the model is falsely labeled as negative. ResNet50+IPM correctly marked the left-upper lobe, and the mass contained in it. ResNet50+WPM only weakly marked this region. No model correctly marked the mass at the left lower lobe. From the example of results shown in Fig. 10, it can be concluded that the trained model failed to learn a mass feature. It also emphasizes the need for more effective feature extraction if we want to detect masses in CXR more accurately.

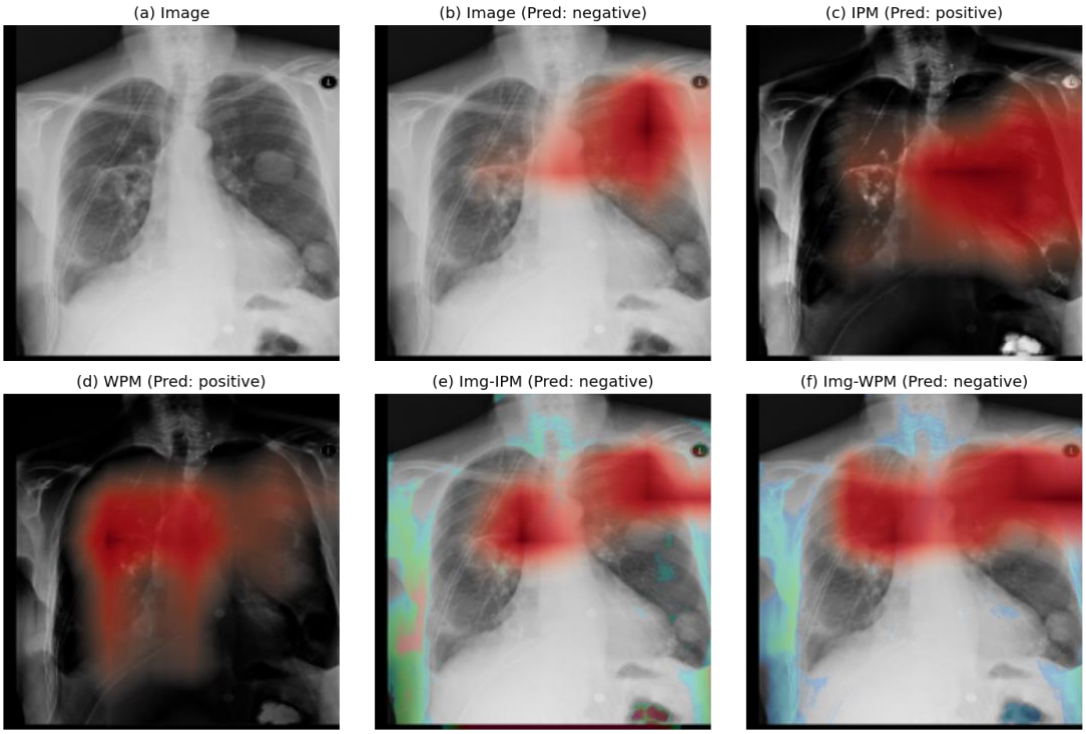

Fig. 10: Output of the Grad-CAM algorithm for the mass/nodule dataset. Regions that are marked red are regions that the ResNet50 deems important.

5. CONCLUSION

In this paper, we introduce the Xenafas algorithm, which creates the IMP and WMP anomalous saliency mapping for CXR images. A qualitative study by a certified radiologist has shown that the algorithm can highlight most foreign objects and cardiomegaly in the CXR samples tested; however, it is inconsistent in highlighting masses and nodules. It has also been shown that using IMP and WMP over regular CXR images increases the sensitivity of most CNN models that were tested. Using the Grad-CAM algorithm, it has been demonstrated that by using the IMP and WMP, the CNN model shifted its focus to a

more relevant CXR image region. The results obtained from the experiment conducted show that the IMP and WMP can be an alternative to regular CXR images for future machine learning development.

ACKNOWLEDGEMENT

This work was supported by the Malaysian Ministry of Higher Education Fundamental Research Grant Scheme [grant no. FRGS17-040-0606].

REFERENCES

[1] Itti L, Koch C, Niebur E. (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell. 20: 1254-1259.

[2] Liu N. and Han J. (2016) Dhsnet: Deep hierarchical saliency network for salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; pp 678-686.

[3] Rahtu E, Kannala J, Salo M, Heikkilä J. (2010) Segmenting salient objects from images and videos. In European Conference on Computer Vision; pp 366-379.

[4] Itti L, Koch C. (2001). Computational modelling of visual attention. Nature Reviews Neuroscience, 2: 194-203.

[5] Cheng M-M, Zhang Z, Lin W-Y, Torr P. (2014) BING: Binarized normed gradients for objectness estimation at 300fps. In Proceedings of the IEEE conference on computer vision and pattern recognition; pp 3286-3293.

[6] Montabone S, Soto A. (2010) Human detection using a mobile platform and novel features derived from a visual saliency mechanism. Image Vis Comput., 28: 391-402

[7] Hou X, Zhang L. (2007) Saliency detection: A spectral residual approach. In IEEE Conference on Computer vision and Pattern Recognition; pp 1-8.

[8] Borji A, Cheng M-M, Hou Q, Jiang H, Li J. (2019) Salient object detection: A survey. Comput Vis Media; pp 1-34

[9] Castillo JC, Tong Y, Zhao J, Zhu F RSNA Bone-age detection using transfer learning and attention mapping [http://noiselab.ucsd.edu/ECE228_2018/Reports/Report6.pdf]

[10] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz CP, others. (2018) Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med., 15: e1002686

[11] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. (2016) Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; pp 2921o2929.

[12] Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. (2019) Deep learning to improve breast cancer detection on screening mammography. Sci. Rep., 9: 12495.

[13] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. (2015) Striving for simplicity: The all convolutional, arXiv preprint arXiv:1412.6806

[14] Ding Y, Sohn JH, Kawczynski MG, Trivedi H, Harnish R, Jenkins NW, Lituiev D, Copeland TP, Aboian MS, Mari Aparici C, others. (2019) A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. Radiology, 290: 456o464

[15] Norman B, Pedoia V, Noworolski A, Link TM, Majumdar S. (2019) Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. J. Digit. Imaging, 32: 471-477

[16] Oh K, Kim W, Shen G, Piao Y, Kang N-I, Oh I-S, Chung YC. (2019) Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization. Schizophr Res., 212: 186-195

[17] Arun NT, Gaw N, Singh P, Chang K, Hoebel KV, Patel J, Gidwani M, Kalpathy-Cramer J. (2020) Assessing the validity of saliency maps for abnormality localization in medical imaging, arXiv preprint arXiv:200600063 Cs

[18] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision; pp 618-626.
[19] Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D. (2019) Attention gated networks: Learning to leverage salient regions in medical images, arXiv preprint arXiv:180808114 Cs
[20] Pesce E, Withey SJ, Ypsilantis P-P, Bakewell R, Goh V, Montana G. (2019) Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. Med Image Anal, 53: 26-38
[21] Deeba F, Bui FM, Wahid KA. (2020) Computer-aided polyp detection based on image enhancement and saliency-based selection. Biomed Signal Process Control, 55: 101530.
[22] Fan H, Xie F, Li Y, Jiang Z, Liu J. (2017) Automatic segmentation of dermoscopy images using saliency combined with Otsu threshold. Comput Biol Med, 85: 75-85
[23] Khan MA, Akram T, Sharif M, Saba T, Javed K, Lali IU, Tanik UJ, Rehman A. (2019) Construction of saliency map and hybrid set of features for efficient segmentation and classification of skin lesion. Microsc Res Tech., 82: 741-763.
[24] Rahmat T, Ismail A, Aliman S. (2018) Chest x-rays image classification in medical image analysis. Appl Med Inform., 40: 63-73.
[25] Otsu N. (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern, 9: 62-66.
[26] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:170404861
[27] Huang G, Liu Z, Weinberger KQ. (2016) Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; pp 4700-4708
[28] He K, Zhang X, Ren S, Sun J. (2016) Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; pp 770-778
[29] Simonyan K, Zisserman A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556
[30] Chollet F. (2016) Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; pp 1251-1258
[31] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; pp 248-255.
[32] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. (2017) Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; pp 2097-2106
[33] Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, Eswaran K, Cameron Chen P-H, Liu Y, Kalidindi SR, et al. (2020) Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology, 294: 421-431