Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(6): 10153–10173.
DOI: 10.3934/mbe.2023445
Received: 11 January 2023
Revised: 12 March 2023
Accepted: 16 March 2023
Published: 29 March 2023

Research article

GL-FusionNet: Fusing global and local features to classify deep and superficial partial thickness burn

Zhiwei Li[1,†], Jie Huang[2,†], Xirui Tong[2], Chenbei Zhang[1], Jianyu Lu[2], Wei Zhang[2], Anping Song[1,*] and Shizhao Ji[2,*]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China

[2] Department of Burn Surgery, the First Affiliated Hospital of Naval Medical University, Shanghai 200444, China

† The authors contributed equally to this work.

* **Correspondence:** Email: Apsong@shu.edu.cn, shizhaoji@aliyun.com.

Abstract: Burns constitute one of the most common injuries in the world, and they can be very painful for the patient. Especially in the judgment of superficial partial thickness burns and deep partial thickness burns, many inexperienced clinicians are easily confused. Therefore, in order to make burn depth classification automated as well as accurate, we have introduced the deep learning method. This methodology uses a U-Net to segment burn wounds. On this basis, a new thickness burn classification model that fuses global and local features (GL-FusionNet) is proposed. For the thickness burn classification model, we use a ResNet50 to extract local features, use a ResNet101 to extract global features, and finally implement the add method to perform feature fusion and obtain the deep partial or superficial partial thickness burn classification results. Burns images are collected clinically, and they are segmented and labeled by professional physicians. Among the segmentation methods, the U-Net used achieved a Dice score of 85.352 and IoU score of 83.916, which are the best results among all of the comparative experiments. In the classification model, different existing classification networks are mainly used, as well as a fusion strategy and feature extraction method that are adjusted to conduct experiments; the proposed fusion network model also achieved the best results. Our method yielded the following: accuracy of 93.523, recall of 93.67, precision of 93.51, and F1-score of 93.513. In addition, the proposed method can quickly complete the auxiliary diagnosis of the wound in the clinic, which can greatly improve the efficiency of the initial diagnosis of burns and the nursing care of clinical medical staff.

Keywords: burn depth; deep learning; partial thickness burns; feature fusion; image classification

1. Introduction

Burn injuries are among the most devastating of all injuries and a major global public health crisis that can be deadly or cause a victim to suffer extremely if not treated appropriately [1]. Catastrophic burn injuries constitute extremely distressing and physically devastating type of trauma that impact approximately every major organ [2]. The World Health Organization estimates that 180,000 people die from burns each year, and in 2004, approximately 11 million people worldwide suffered from severe burns that required treatment [3]. Radiation, electricity, heat, excessive cold, chemical elements, etc., can cause severe burn injuries, and treatments must be ensured carefully according to its severity [4]. The survival rate of burn patients can be greatly improved with early and adequate treatment. In the early stages of burn wound treatment, excision, skin grafting and skin replacement are typical treatment techniques. Through these approaches, the outcomes of severely burned patients can be improved by reducing mortality and length of hospital stay. And if the right treatment is not done at the right time, it can have serious consequences. For example, poor wound healing, infection, discomfort, hypertrophic scarring, organ failure, and even death may ensue [5].

While superficial and full thickness burns are straightforward to diagnose based on visual appearance, clinicians have difficulty with accurate differentiation between superficial partial and deep partial thickness burns [6]. The experience of the clinician and the fact that these burns can dynamically increase in severity (i.e., burn wound conversion) during the initial 48-h period lead to higher clinician error [7]. Both wound types are associated with similar characteristics, and both involve the epidermis and dermis. Superficial partial-thickness burns involve injury of the papillary dermis and are associated with intact blisters, moderate edema, a moist surface under the blisters, a bright pink or red color, and blanching with a fast capillary refill after pressure is applied. Deep partial-thickness burns involve injury of both the papillary and reticular dermis and are associated with broken blisters, substantial edema, a wet surface, a mixed red or waxy white color, and blanching with a slow capillary refill after pressure is applied [8]. These problems will make clinicians misjudge when judging the depth of burns, but there is currently no effective method to quickly assist physicians in making accurate diagnoses. With the widespread use of computer methods in the medical field in recent years [9], we see the possibility of applying computer methods to this classification task.

However, many current judgment methods require more equipment and technologies, such as burn wound biopsy [10], hyperspectral imagery [11], ultrasound imagery [12] and polarized light photography images [13], which have been used for artificial intelligence models of burn depth.

Aiming at some addressing problems and the current research status in the field of burn diagnosis, this paper mainly focuses on the following three aspects:

- First-time use of a deep learning method that focuses on the deep and superficial classification of partial-thickness burns to aid diagnosis.
- According to the characteristics of burn depth classification, a fusion model combined with a segmentation network is proposed to improve the classification accuracy of deep and superficial partial-thickness burns.
- It is of practical significance to conduct experiments on the dataset collected clinically, so the proposed method is easy to practice clinically.

2. Related works

In recent years, computer technology has been continuously applied in the medical domain, especially in clinical practice, which has a very good auxiliary effect in terms of improving the clinical diagnosis ability of doctors [14]. And all over the world, there are studies that have used various computing techniques to automatically classify burn images, and judge the severity of burn injuries in real time based on the captured damage images [15]. In the field of image classification, machine learning methods are one of the most widely used techniques. Generally speaking, machine learning methods analyze and retrieve key information from a large amount of heterogeneous data, and then they use this information to autonomously detect and classify different categories [16]. Therefore, employing various machine learning techniques for burn severity assessment is gaining traction nowadays. For example, the work [17] had proposed a method for categorizing burn photos into the second, third, and fourth degrees of severity, in which they used a combination of image processing techniques concentrating on color feature extraction from the images and then support vector machine (SVM) classifiers to categorize the images. And, a group [18] used 105 burnt photos to develop an automatic segmentation-based classification method to categorize burn images into healthy skin, burned skin, and background, for which they employed four types of clustering approaches for image segmentation and then applied several traditional machine learning classification techniques with an aim of exploring the best-performing classifier. In the study of [19], the authors utilized 74 burn images to develop a feature extraction model with several digital image processing steps; they then classified the images into two classes using a SVM classifier. Another group [20] proposed a real-time technique for classification of burn depth employing moderate sample sizes based on ultrasound images. They constructed the textural feature set by using a grey-level co-occurrence matrix derived from the ultrasound images of the burn tissue; they then utilized a nonlinear SVM and kernel Fisher discriminant analysis. Finally, they completed the validation of the classification effect by using pig skin tissue under four different burn scenarios.

With the increasing popularity of deep learning in recent years, some studies have also applied deep learning technology in the fields of burn classification and severity detection. For example, the authors developed a deep learning-based system in [21], and it included precise burn area segmentation and burn depth labeling; they also proposed a framework for enhanced burn area segmentation and automated burn depth diagnosis based on deep learning methods. And, the authors of [22] proposed a predictive model based on a deep neural network, recurrent neural network(RNN) and convolutional neural network (CNN) to determine degree 1, degree 2 and degree 3 of burn images depending on the severity of the burn for a dataset of 104 images. In another study [23], the authors presented a DenseMask regional CNN technique, which combined a mask-region based convolution neural network with dense pose estimation for segmenting the region of interest of a skin burn area from images based on the severity of the burn damage. Another work proposed in [24] applied a deep n eural network with transfer learning using two pre-trained models ResNet50 and VGG16 for the feature extraction from images; the method then applied an SVM classification approach to classify the images into four categories, i.e., healthy skin, first-degree, second-degree and third-degree burns on 2080 RGB input images. Also, the authors of [25] proposed a novel method that employs a state-of-the-art deep learning technique to segment the burn wounds in the images. They designed this deep learning segmentation framework based on the mask regions with a convolutional neural

network (Mask R-CNN) and obtained very good results.

However, it can be found that the machine learning and deep learning methods mentioned above focus on classifying burns in four degrees. At present, there are few studies that have focused on deep partial-thickness and superficial partial-thickness burns. At the same time, although the existing studies have segmented local wounds, they lack a method to connect local and global features. Therefore, we propose the GL-FusionNet to try to do some research work in response to these problems.

3. Method

In this paper, we summarize some of the features of previous studies on trauma segmentation and classification. Based on the burn trauma segmentation results obtained using the segmentation network of U-Net, we carried out research related to the classification and area measurement of deep and superficial partial-thickness burns. In the classification study, to obtain more accurate results, we used an ResNet101 and ResNet50 to extract the global and local features of the original image and segmentation results respectively and then weighted the two features to superimpose them; we finally used the fused features for classification, achieving better results.

3.1. Network architecture

As shown in Figure 1, the network structure of our main classification network consists of two stages. The first stage is to use the segmentation network to segment the original input burn wound image to obtain a segmentation mask. The segmentation mask is then superimposed on the original image to obtain an image containing only the wound portion (referred to as the segmentation result image in the later section). The second stage involves using our proposed classification network to perform the classification of deep and superficial partial-thickness burn depths. Both the original image and the segmentation result image obtained in the first step were used as input to the second stage, where we considered the features extracted directly from the original image as global features and the segmentation result image as local features. The global and local features are then combined and fed into the classifier to obtain the final classification result. Of course, at each input and output of the network, we have adapted the structure of the input and output to suit our dataset.

For each of the extracted features in segmentation and classification, we tried different models to achieve the best results. ResNet101 was used to extract the global features, while ResNet50 was used to extract the local features from the segmented images. The features extracted from both were eventually superimposed using certain fusion strategies, resulting in a fused feature of the original image (combining global and local).

Finally, the resulting fused features (in Figure 1) were compressed by using a fully connected layer, and the output was a two-category compressed feature. The resulting features were then transformed into the output using the softmax function (two output classes).

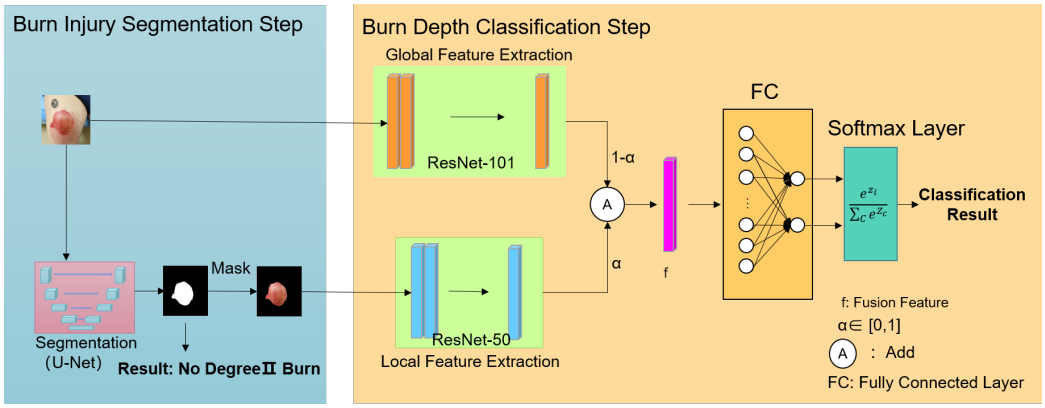

Figure 1. Framework of the burn depth classification network. The framework comprises two steps. The first step is the burn injury segmentation stage, and the second step is the burn depth classification stage.

3.2. Two stages and three classifications

The classification model framework proposed in this paper can be divided into two main stages. The first stage is to segment the original image (for partial-thickness burn trauma) to obtain the exact location of the partial-thickness burns in the original image. Also, output is performed at the end of this stage indicating whether the original image is a partial-thickness burn, and if it is not, the algorithm is directly terminated; if it is a partial-thickness burn, it proceeds to the next step of the categorization and detection process. It should be noted that in the process of discriminating whether it is a partial-thickness burn or not, the effect of mis-segmentation is excluded by setting a threshold value.

Throughout the first stage, it is the U-Net that is most predominantly used for the segmentation of burn trauma. Traditional image segmentation algorithms generally use certain fixed common features, such as color features, texture features, edge features, etc. Such methods have high interpretability but low accuracy, and they often cannot complete image segmentation tasks well. So, we chose to use the U-net of deep neural networks for the first stage of the segmentation task. Compared with other segmentation networks with deep learning (FCN [26], Mask R-CNN, etc.), the biggest difference with U-Net [27] is the choice of stitching for feature fusion instead of point-to-point superposition, which forms thicker features and makes the final features richer. Combined with the multi-level feature extraction framework, U-Net can get accurate classification results from a small number of training images. Therefore, the U-Net has a good application scenario in the field of medical images where data and annotation are difficult to obtain. As shown in Figure 1, the overall framework of the U-Net has a simple structure. The left half uses a shrinkage path for feature extraction, and the right half is upsampled by an expansion path to finally obtain the segmentation result image. In our model, the input and output are adjusted to make them suitable for our dataset. In addition, some researchers have also applied deep learning-based image segmentation networks in many medical fields. But, the main research objects of previous studies [28–31] were professional CT images, while our research objects are general images collected by ordinary equipment, and the final application also needs to be applied to general color images. At the same time, the improved network models used in other studies [21–24]

are more complex, and the generalization has not been widely verified. Therefore, instead of using these improved networks, we use the classic U-Net for the first stage of wound segmentation.

And, at the end of the whole classification process, it will output whether the input image is a deep partial-thickness or shallow partial-thickness burn trauma according to the classification result. In other words, the classification framework designed in this study is a sub-classification framework, in which the depth of burns is first roughly classified in two frameworks to obtain partial-thickness burns, outputting other depths of non-second-degree burns, and further classifying superficial partial-thickness and deep partial-thickness burns of partial-thickness burn depth in the second stage. It is hoped that the accuracy of the final classification results can be further improved by such a method.

3.3. Local feature and global feature

Of the existing studies that address the classification of burn depth (the results of all burn classifications are combined because there are fewer studies that focus only on the classification within partial-thickness burns), two main types of datasets are used; one type is shown in Figure 1 and it uses the complete picture for the study (containing some degree of environment and surrounding skin), and the other type segments the original picture and uses the segmented picture that contains only the wound part of the image for the study. The former can be said to use the global information for feature extraction, so that product can be called global features, while the latter contains only the information of the wound part, so that product can be called local features.

Global features, as a comprehensive feature, are rich in information and can describe a wound as a whole. However, global features cannot contain spatial information and cannot know which part of the global space is more important. At the same time, global features are easily disturbed by noisy regions, and some background and cluttered information can have a relatively large impact on the final result. Unlike the general local features, the local features defined in this study are also for a part of the region, but the specific improved method is to process the original image first so that the image contains only the wound part, and then to extract features for the image; then, the extracted information is the local features, and the rest of the location is set as blank (no information). However, the local features lose a lot of extra information because there is no remaining part.

The clinicians from Changhai Hospital in our team concluded based on their diagnostic experience that clinicians should not only pay attention to the local wound surface, but should also understand the morphological basis for judging the depth of burn wounds from images. This includes the skin, as well as the surrounding area, which is also informative for the final depth judgment. Therefore, combining some advantages of existing approaches, we propose the approach of wanting to combine global features and local features to obtain highly accurate classification results.

The ResNet was proposed in 2016 by He et al. [32]. In the CNN, the deeper the network, the more features it can obtain. Although this feature of the CNN leads to a series of breakthroughs in image classification, it also brings a lot of new problems. For example, as the network depth increases, the notorious problems of vanishing/exploding gradient occur. The deep residual learning network, which adds a reference at each layer to learn the residual function, can address the degradation problem properly.

Since global information tends to contain more information, ResNet101, which has a deeper network depth, is used for feature extraction in our framework. Since only the information of some regions needs to be extracted, a relatively shallow ResNet50 is used to extract local features.

3.4. Feature fusion strategy

In the application of deep learning, the feature fusion strategy is actually a very common strategy. The feature fusion method we adopted belongs to the method of early fusion, that is, the feature fusion is performed in the stage of feature extraction; the feature fusion is performed first, and then the classifier is trained on the fusion features. This method mainly contains two fusion methods for the feature fusion strategy, where one is the concat method and the other is the add method.

The concat method directly connects multiple features end-to-end. If the dimensions of the two input features are p and q, the dimension of the resulting fusion feature is $p + q$. This method can fuse the information in all channels as a whole through the use of the convolutional kernel, which can enhance the overall classification accuracy. But, this can lead to jumbled information in multiple channels, which can be confusing. The add method directly superimposes the values of the two feature vectors at each channel, and the final fusion feature dimension is the same as the original input. This method needs to ensure that the feature dimensions of the two inputs are the same, and the semantic features of the corresponding channels are similar; otherwise, it will not work. It can ensure that the information at the corresponding channel is fused to enhance the effect.

In our study, we adopted the fusion strategy of the add method. In our framework, the features after two-way feature extraction can be adjusted one-to-one. At the same time, the extraction of local features and global features comes from the same original image, the information at the corresponding channels is consistent and the use of the add method can effectively avoid the confusion of semantic information between different channels.

In addition, since the extraction of local features in our framework comes from the original image, the fusion strategy of the add method is actually equivalent to making the more important parts of the local features more prominent in the global features. By analogy, it is a method similar to the attention module. By increasing the importance and proportion of local features in the global features, the network model can pay more attention to the wound information extracted by the local features, and at the same time, it does not give up the wide area in the global features. This kind of method can effectively improve the accuracy of the final classification.

Finally, our fusion method differs from the common add method by adding weighted coefficients. According to the diagnostic experience of clinicians on our team, for the judgment of burn depth, the information about the wound area is more important than the information about the surrounding area and other areas. Therefore, based on the experience of clinicians, we propose three weighting ratios of 5:5, 4:6 and 3:7 for global features and local features. After the final experiment, a weighted ratio of 4:6 was used as the final fusion strategy.

3.5. Loss function

In our classification model framework, the calculation of the loss function mainly involves the loss of fused features. The final output targets the classification results, and we use a feature fusion approach; so, instead of calculating the loss of the two feature extraction branches, the loss of the fused features is calculated uniformly.

And, we use cross-entropy loss as a loss function. The cross-entropy loss function is adopted to convert the output of the network (logits) into probabilities using the softmax function. The output

probability can be calculated by using Eq (3.1):

$$p_j^{(i)} = \text{Softmax}(z_j^{(i)}) = \frac{\exp(z_j^{(i)})}{\sum_{m=0}^{M} \exp(z_j^{(m)})} \tag{3.1}$$

where $z_j^{(i)}$ and $p_j^{(i)}$ denote the logit and probability values of the jth speech of the ith class, respectively, and M is the number of classes in the training set. Thus the cross-entropy loss function is given by Eq (3.2):

$$\mathcal{L} = -\frac{1}{B} \sum_{i=0}^{B} \sum_{c=0}^{C} y_i^{(c)} \log p_i^{(c)} \tag{3.2}$$

where $y_i^{(c)}$ is the ground truth of the ith sample of the cth class, C is the number of classes and B is the value of the batch size during training. The derivation of cross-entropy loss for multiple classes is more straightforward and better at learning inter-class information. The gradient of the last layer is not correlated with the derivative of the activation function, leading to a faster update of the weight matrix and faster convergence during training.

3.6. Training details

In order to get better training results, we do not use random parameters when initializing the parameters. In the initialization phase of the network for segmentation, as well as for classification, we use the migration learning approach. The model weights after pre-training with ImageNet were used as initialization parameters. The same pre-trained weights were also used as initialization parameters to ensure the validity of the comparison experiments. Finally, the network model and weights were fine-tuned by collecting data using a migration learning approach.

In addition, for the comparison of the classification models, we used approximately the same training environment and parameter settings in order to fairly judge the results of the comparison experiments and the original experiments. As shown in Table 1, the number of training epochs was 600, the batch size was 64, the optimizer was SGD, the initial learning rate was set to 1×10^{-3} and the minimum learning rate was 1×10^{-5}. The learning rate is automatically adjusted according to the number of training epochs.

Table 1. Training parameters.

Parameters	**Values**
Epoch	600
Batch Size	64
Optimizer	SGD
Learning Rate	1×10^{-3}
Min Learning Rate	1×10^{-5}

Finally, it should be noted that we directly uses the original images and the corresponding segmentation annotations (called ground truth) from the first stage in the training phase of the classification model in order to make the best training results. That is, a set of corresponding original images and ground truth are used as the input of the classification stage, and one final output of the classification results will be obtained.

4. Experimental Result

The results of the experiments in this paper are divided into two main sections, including the results of the segmentation experiments and the results of the classification experiments. Each part of the experiment also included a comparison experiment and an ablation experiment.

4.1. Dataset

Starting in 2021, we began collaborating with the Burn Department of Shanghai Changhai Hospital on the collection and labeling of data. The study was approved by the Shanghai Changhai Hospital Ethics Committee, and all methods were performed in accordance with the relevant guidelines and regulations. Written informed consent was obtained from all patients for sample collection. During the data collection phase, nurses in the burn unit at Changhai Hospital used smartphones to collect images of patients' burns. At the same time, all images were captured with the same equipment. The equipment model was Honor Magic 2, the focal length was set to 27 mm and the imaging size is 3456×4608. Thinking of standardization of data collection, we have established a number of rules for capturing burn images, including selecting deep and superficial partial-thickness burn wounds to photograph; photographing vertically directly above the wound; photographing under normal lighting conditions (not overexposed); and photographing to avoid gauze, blood, etc., in the image. For professionalism and accuracy in annotation, postgraduate students from the Naval Medical University, under the guidance of nurses and physicians, used Lableme (software used for annotation) to form accurate annotations, as shown in Figure 2, including a delineation of burn expectancy and a classification of burn depth results. Ultimately, we collated and annotated 500 original partial-thickness burn images, 268 deep partial-thickness burn images and 278 superficial partial-thickness burn images. Furthermore, it is important to note that the images chosen contain only one depth of the wound. Finally, for training convenience, all images were resized to 512×512.

Figure 2. Annotation tool with annotations including segmented annotations and the depth of wounds. Annotation by several professionals using Labelme, an open-source annotation tool.

Due to the difficulties of data acquisition and the lack of data volume, we used data augmentation methods, including the generation of slightly modified copies of images from the original training samples, as an effective strategy to reduce data scarcity, improve performance and to minimize prediction errors. We used three methods based on geometric transformations, including flipping, random rotation and random cropping. The final augmented dataset of 3264 images had been generated by applying augmentation techniques.

At the same time, it should be noted that, in order to ensure the reliability of the experiment, all experiments in this study were carried out using a 50-fold cross-validation method. First, the dataset was evenly divided into five parts (to ensure that the number of each classification in each part was roughly the same). Then we took one of them as the test set each time without repeating, used the other four parts as the training set to train the model, and verified it on the taken-out test set. Finally, we calculated the average value of five training and verification times as the final result. This approach can yield a good effect on the small dataset used in this study and it has a good verification effect in terms of the generalization and accuracy of the model.

4.2. Metrics

For the classification model, several common metrics were selected to measure the accuracy and generalization of the model in combination. The classification process of a dataset produces four categories of detection results TP (true positive), FP (false positive), FN (false negative) and TN (true negative). The four evaluation metrics selected therein are obtained by using arithmetic (all metrics selected are of higher value, representing better model effectiveness). The accuracy rate, as shown in Eq (4.1), is the most intuitive indicator, directly calculating the ratio of correct data to total data.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

Precision, as shown in Eq (4.2), represents the proportion of true positive samples in the precision results, and it allows you to assess how well the classifier has been able to classify the data based on success.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Recall, as shown in Eq (4.3), is the proportion of all positive cases that are correctly predicted, and it is used to assess the coverage of the classifier over all samples classified.

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

Where, since precision and recall cannot be high in most cases, the F1-score is used to combine precision and recall, as shown in Eq (4.4). and it is the harmonic mean of precision and recall.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.4}$$

Due to the difference between segmentation and classification, all segmentation experiments in this study use the Dice score [33] and Intersection-over-Union (IoU) score, which are commonly used in

segmentation experiments, to evaluate the model's performance. The calculation formulas of the Dice score and IoU score are respectively shown in Eqs (4.5) and (4.6).

$$Dice = \frac{2 \mid X \cap Y \mid}{\mid X \mid + \mid Y \mid} \tag{4.5}$$

$$IoU = \frac{\mid X \cap Y \mid}{\mid X \mid + \mid Y \mid - \mid X \cap Y \mid} \tag{4.6}$$

In detail, X is the predicted result, Y is the actual result, $X \in [0, 1]$, $Y \in [0, 1]$, $\cap$ is the intersection between the two, which can be approximated as the dot product between the prediction map and the ground truth, $Dice \in [0, 1]$ and $IoU \in [0, 1]$.

4.3. Segmentation results

Common segmentation models are used as a comparison to conduct experiments. The specific segmentation experimental results are shown in Table 2. It can be seen that the segmentation model using U-Net has achieved the highest scores in both *Dice* and *IoU* on the wound dataset, and this result also verifies the correctness of choosing U-Net as the segmentation model. It should be noted that the segmentation tasks here only performed the segmentation of second-class burn wounds, and did not directly perform multi-classification of superficial and deep partial-thickness burns. In comparative experiments, we used the U-Net, Res-UNet, DeepLabV3 with MobileNet, and U-Net with MobileNet. At the same time, we implemented the FP-U-Net++ [31] for testing, and the effect on our dataset was not much different from the general U-Net results. This is why we finally chose to use a U-Net with a simpler network model.

Table 2. Segmentation result. Bold entries indicate the best-performing experiments.

Method	**Dice (%)**	**IoU (%)**
U-Net	**92.587**	88.726
Residual U-Net	92.413	86.782
U-Net (MobileNetV2)	77.875	72.856
DeepLabV3+ (MobileNetV2)	87.632	84.583
FP-U-Net++	92.451	**88.863**

Also, as shown in Figure 3, we compared the effects of various segmented networks. It can be said that the U-Net we used has the best segmentation results on wounds of various sizes.

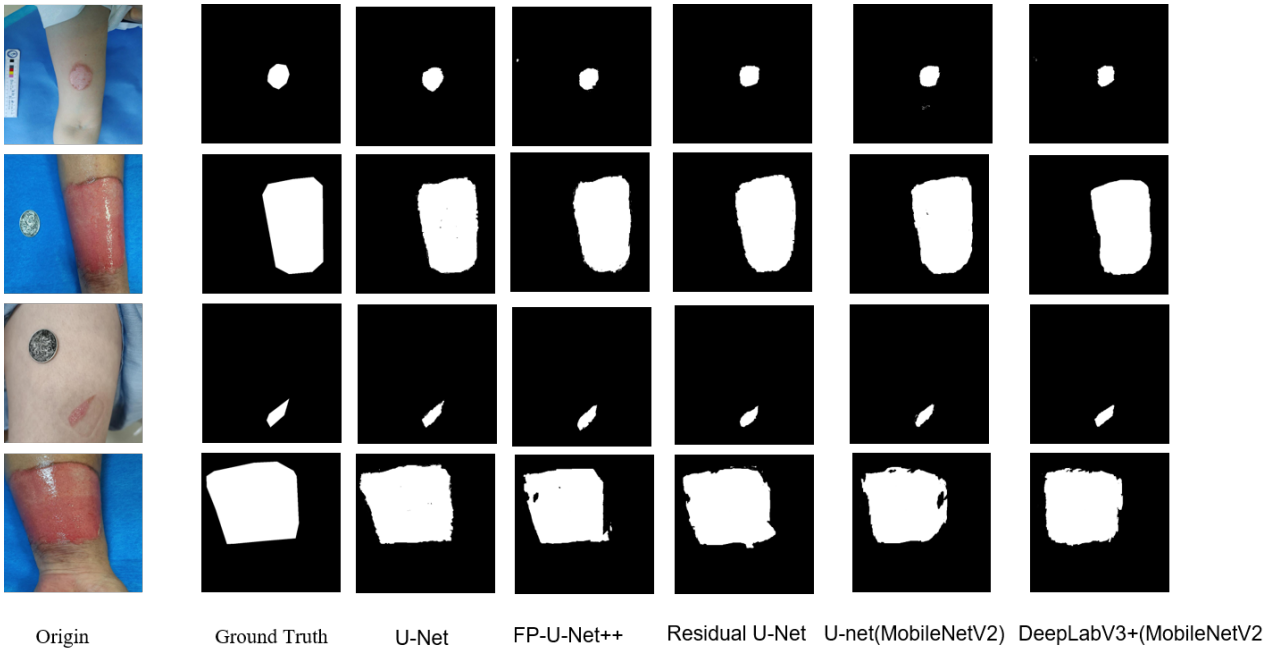

Figure 3. Performance of different segmentation methods.

Finally, the change in the number of training loss iterations is shown in Figure 4. In general, the network converges at a low loss level after training, and it can be said that the obtained model has been fully trained.

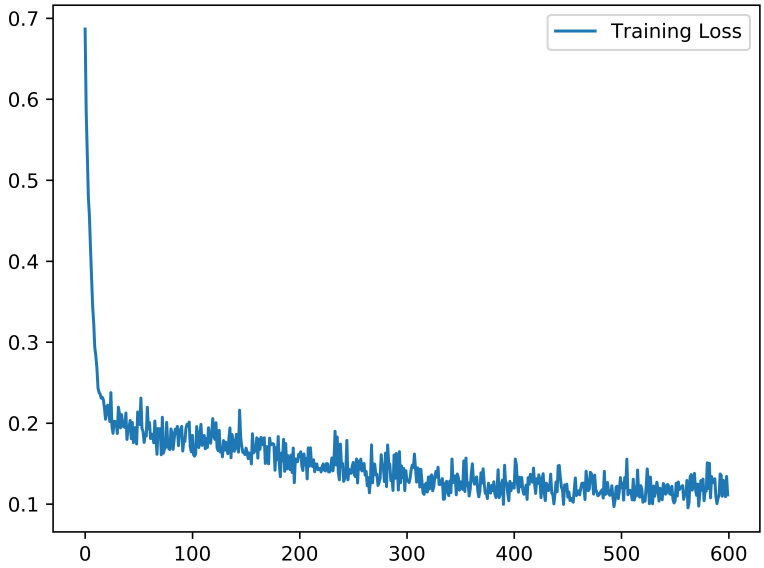

Figure 4. Training time and convergence curve with iterations for U-Net training. The x axis represents the number of iterations and the y axis represents the value of loss.

4.4. Multi-class segmentation results

The current segmentation model can also directly perform multi-classification tasks in addition to binary classification (superficial and deep partial-thickness burns), so experiments with multi-classification segmentation were also conducted to verify the segmentation results. The results are shown in Table 3, and it can be seen that, although the segmentation model using the U-Net still achieved the highest scores in terms of average Dice and IoU, there was a significant decrease in overall accuracy relative to the results of the binary classification segmentation. It can be seen that, although the multi-class segmentation can directly achieve the ultimate goal, it is not effective enough to achieve a certain level of classification. It should be noted that the network model that performed better in the binary segmentation experiments was selected for this part of experiment (the U-net (MobileNetV2) network model, which did not work as well, was dropped).

Table 3. Multi-class segmentation result. Bold entries indicate the best-performing experiments.

Method	**Dice (%)**	**IoU (%)**
U-net	**85.352**	**83.916**
Residual U-Net	77.167	76.244
DeepLabV3+(MobileNetV2)	79.195	75.354

4.5. Classification results

For the classification experiments, we designed three types of comparative experiments, including using only global features, using only local features and using our fusion model. Since the results of the three types of experiments have the same meaning, we display the three types of experiments in the same table for clear presentation. In this part, we used ResNet34,ResNet50, ResNet101, ResNet152, VGG16, VGG19, EfficientNet-B0 and EfficientNet-B7.

As shown in Table 4, it can be seen that our model using fused features achieved the best value for each evaluation metric among all of the models. At the same time, it can be found that the results of the model using only local features were overall better than those of the model using only global features. In the model that uses local features, the model with deeper network layers will get better results. In contrast, in the model that only uses global features, the model with more shallow network layers will get better results. This also verifies the accuracy of our approach of using deep networks to extract global features and shallow networks to extract local features.

Finally, to further measure the generalization of our method, we used the receiver operating characteristic (ROC) and area Under the receiver operating characteristic curve (AUC) for evaluation. As shown in Figure 5, our method achieved the best result on the AUC, reaching 0.97. At the same time, it can be seen that methods using global features generalize more than methods using local features. This shows that the global features may play a certain positive role in judging the classification of the deep and superficial partial-thickness of depth.

Table 4. Comparison of five-fold classification results of different models. The results of accuracy, recall, precision and F1-score are present in this table. The results are the combined results of five-fold cross-validation. We show the MEAN±STD (standard deviation) scores of five trained models of each training validation fold. Bold entries indicate the best-performing experiments.

Methods	Metrics			
	Accuracy	Recall	Precision	F1-Score
Fusion (ResNet101+ResNet50)-ours	**93.523±1.451**	**93.670±1.374**	**93.510±1.459**	**93.513±1.458**
ResNet101 (Global)	85.148±2.893	82.078±2,721	82.334±1.558	81.704±1.648
ResNet50 (Global)	73.226±1.550	73.468±2.177	74.094±1.301	72.904±1.768
ResNet152 (Global)	82.500±3.446	82.610±3.486	82.270±3.446	82.328±3.444
ResNet34 (Global)	70.124±2.224	69.874±2.405	70.058±2.218	69.796±2.218
VGG16 (Global)	76.670±2.475	76.522±2.640	76.754±2.395	76.542±2.477
VGG19 (Global)	81.602±2.652	81.780±2.435	82.034±2.198	81.548±2.661
EfficientNet-B0 (Global)	69.870±4.830	69.790±4.958	70.122±4.604	69.526±5.085
EfficientNet-B7 (Global)	83.130±2.076	83.324±2.102	83.154±2.165	82.728±1.761
ResNet101 (Local)	89.053±0.796	87.467±2.391	84.703±4.095	86.976±2.836
ResNet50 (Local)	91.650±1.542	91.660±1.518	91.713±1.491	91.596±1.571
ResNet152 (Local)	74.164±2.411	74.268±2.598	74.596±2.676	73.982±2.492
ResNet34 (Local)	79.026±2.173	79.086±2.204	79.034±2.164	78.924±2.234
VGG16 (Local)	80.305±3.808	80.637±4.113	80.222±3.742	80.201±3.764
VGG19 (Local)	76.474±1.912	76.608±1.850	76.654±1.691	76.442±1.898
EfficientNet-B0 (Local)	57.842±2.286	57.996±1.909	58.154±1.892	57.588±2.125
EfficientNet-B7 (Local)	68.840±2.891	68.772±2.909	68.928±3.064	68.821±3.036

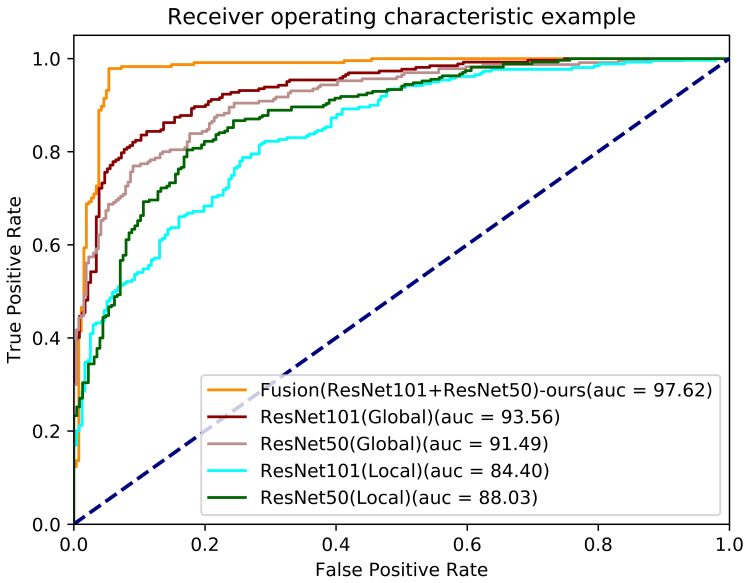

Figure 5. Some extension of ROC. The AUC is also marked in the legend.

At the same time, the changes in the number of training loss iterations are shown in Figure 6. After training, the network convergence is at a low level of loss, and it can be said that the obtained model has been fully trained. But, what can be seen is that the loss is high at the beginning, and the final convergence change is not easy to observe.

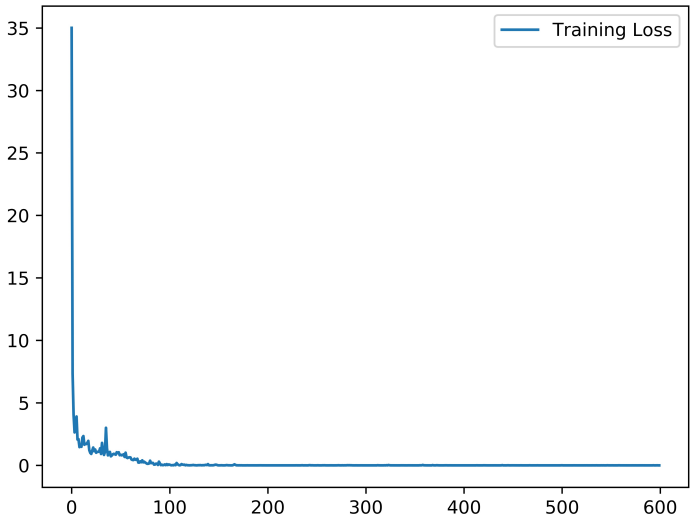

Figure 6. Training time and convergence curve with iterations for U-Net training. The x axis represents the number of iterations and the y axis represents the value of loss.

4.6. Comparison experiments results

For our fusion feature model, there are two points that need to be verified by comparative experiments. The first is the selection of the feature extraction network for the two types of features. We selected the network models with better effects in the separate feature models for combination. The experimental results are shown in Table 5, and it can be seen that our fusion method achieved the best results for each comparative experiment.

Table 5. Comparison of five-fold classification results for our GL-FusionNet with different branched network structures. The results of accuracy, recall, precision and F1-score are present in this table. The results are the combined results of five-fold cross-validation. We show the MEAN±STD scores of five trained models of each training validation fold. Bold entries indicate the best-performing experiments.

Methods	Metrics			
	Accuracy	F1-Score	Precision	Recall
Fusion (ResNet101+ResNet50)	**93.523±1.451**	**93.670±1.374**	**93.510±1.459**	**93.513±1.458**
Fusion (ResNet50+ResNet101)	73.230±3.197	72.133±3.556	73.313±4.439	73.233±3.188
Fusion (ResNet152+ResNet50)	83.386±1.593	83.267±1.563	84.263±1.869	83.346±1.594
Fusion (ResNet50+ResNet152)	61.783±2.591	61.640±2.904	62.316±2.478	61.526±1.992

The other is the choice of weighting coefficients. Based on the judgments and suggestions of clinicians, we verified the models of three weighted-ratio feature fusion methods respectively. The experimental results are shown in Table 6. It can be seen that when the ratio of local to global features is 6:4, the effect of the model is the best, which also verifies the accuracy and effectiveness of our method.

Table 6. Comparison of five-fold classification results for our GL-FusionNet with different weights. The results of accuracy, recall, precision and F1-score are present in this table. The results are the combined results of five-fold cross-validation. We show the MEAN±STD scores of five trained models of each training validation fold. Bold entries indicate the best-performing experiments.

α	Metrics			
	Accuracy	F1-Score	Precision	Recall
0.5	89.472±0.934	89.510±1.237	89.108±0.365	89.076±1.089
0.6	**93.523±1.451**	**92.513±1.458**	**93.510±1.459**	**92.677±1.374**
0.7	90.015±1.530	90.014±1.470	90.005±0.525	89.092±1.539

4.7. Results in different situations

To evaluate the clinical performance of the method proposed in this paper, we additionally collected 100 images of partial-thickness burn wounds collected clinically. The images were classified by specialist physicians on the team (depending on images alone). At the same time, we classified all burn wound images in the validation set according to the cause of injury and gender and evaluated them in different aspects.

As shown in Table 7, among the collected 51 images caused by boiling water, our method performed best. And, in the case of flame (31 images) and other (19 images) types of burns, our method also demonstrated high accuracy. At the same time, it can be seen that the effect of our method achieved good results on burn wounds with various causes. The phenomenon of low classification results for flame and other types of wounds may be due to the insufficient amount of data due to the small number of patients clinically admitted for these two types of wounds. More cases will continue to be collected in the future to obtain more generalization results.

Table 7. Classification performance of our GL-FusionNet on burn wounds with different causes of injury. Bold entries indicate the best-performing experiments.

Cause of burns	Metrics			
	Accuracy	F1-Score	Precision	Recall
Flame	92	88	88	90
Hot water	90.32	87.09	80.64	87.09
Other	89.47	84.21	84.21	78.94

In addition, as shown in Table 8, we also classified the collected validation set by gender and conducted experiments. Across 58 images of men and 42 images of women, our method achieved similar conclusions and relatively high results for various metrics. Combined with the results of various experiments, our method has achieved relatively generalized results in clinical experiments.

Table 8. Classification performance of our GL-FusionNet on burn wounds for different genders.

Gender	Metrics			
	Accuracy	F1-Score	Precision	Recall
Male	89.65	86.21	82.75	84.48
Female	88.09	83.33	85.71	80.95

5. Conclusions and discussion

This study includes two main aspects. The first is the segmentation of partial-thickness burn wounds, and then the further classification of deep and superficial partial-thickness burns based on the segmentation results. Corresponding experiments and studies were carried out with respect to these two aspects, and we achieved certain results.

Wound segmentation, which is the first step in all wound healing and intelligent diagnostic studies, is the basis for many studies to be conducted. In this study, we selected the currently used image segmentation frameworks, performed comparison experiments and selected the optimal U-Net as the final segmentation method. After optimization and adjustment of the model, as well as the fine-tuning of the training, the overall segmentation results were satisfactory enough to serve as a basis for subsequent studies. However, since we did not make major adjustments to the model, there are still many areas that can be improved, such as targeted model modifications for the characteristics of burns with large area variations. Meanwhile, due to the difficulty of dataset collection, this study finally realized only the segmentation model for partial-thickness burn wounds, and no further experiments and comparisons were done for non-partial-thickness burn wounds.

For further classification of partial-thickness burns, we used fusion features incorporating local features and global features to accurately classify superficial and deep partial-thickness burns, and we also conducted comparison experiments to verify the effectiveness of our framework. It also lays a solid foundation for further research in the follow-up. However, since the framework extracts local and global features separately, the classification accuracy is improved, but the speed of detection and the model size are increased due to the relatively large model. One direction of effort for future research is to further optimize the model, reduce the number of parameters and ultimately improve the classification efficiency. Another direction of improvement is to further collect a more complete dataset to extend the research area. The classification of partial-thickness burns can be subsequently extended to the classification of overall burns (mainly due to the difficulty of dataset collection and labeling). In addition, in the process of classifying deep partial-thickness and superficial partial-thickness burns, we found that there are actually large differences in the images of one type of burn classification. For example, among the superficial partial-thickness burn wounds shown in Figure 7, the wound on the right has a tendency to heal, and the overall condition is more different compared to the wound on the left, but both wounds are classified as superficial partial-thickness burn wounds. Therefore, we wound whether the classification of superficial and deep partial-thickness burns could be further classified into more detailed depth classification based on the condition of the wounds. However, this point also needs more argumentation and research, and it is a direction for future work.

In general, our research and experiments on the above two aspects can play a certain positive role

in the existing research fields, and it also achieved ideal results in the experiments. In the future, we will continue to conduct in-depth research, including the use of deep segmentation methods, further sub-classification of second-degree burns and clinical application experiments. Finally, it is hoped that the experimental results can be applied to the clinic and provide an auxiliary role for actual clinical diagnosis.

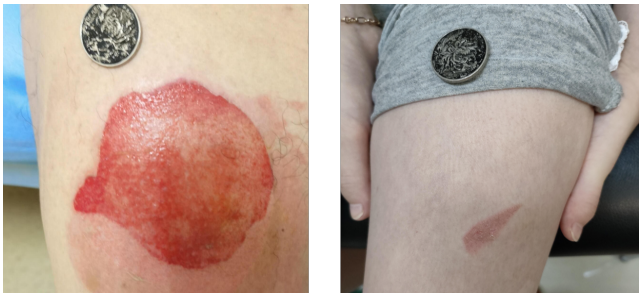

Figure 7. Comparison of two superficial partial-thickness burn wounds.

All in all, in this study, two main experiments were conducted; first, experiments on burn wound segmentation were carried out and a Dice score of 92.587 was obtained on the validation set using the U-Net model. Then, based on this, the classification problem of deep and superficial partial-thickness burns and the study of burn wound area measurement methods were evaluated. In the classification problem, we used deep learning to classify deep and superficial partial-thickness burn injuries for the first time, and we proposed a new model that incorporates global and local features; finally, an average accuracy of 93.52 was obtained on the validation set. Our proposed method is based on deep learning and computer-based automation, which is easier than traditional methods and can greatly improve the efficiency of burn diagnosis and care for clinical staff. It also has great benefits for hospital management and intelligent construction.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (81971836), Deep Blue Talent Project of Naval Medical University and 234 Academic Climbing Programme of Changhai Hospital.

We thank the editor and all of the anonymous reviewers for their valuable advice and suggestions to improve the quality of the current work. We also thank the data collectors for their data collection work. At the same time, we are grateful to the patients who provided their wound to be data.

Conflict of interest

The authors declare that they have no competing interests.

References

1. M. D. Peck, M. Jeschke, K. Collins, Epidemiology of burn injuries globally, *Burns*, **37** (2011), 1087–1274.

2. M. G. Jeschke, G. G. Gauglitz, G. A. Kulp, C. C. Finnerty, F. N. Williams, R. Kraft, et al., Long-term persistance of the pathophysiologic response to severe burn injury, *PLoS One*, **6** (2011), 21245. https://doi.org/10.1371/journal.pone.0021245

3. Y. Wang, J. Beekman, J. Hew, S. Jackson, A. C. Issler-Fisher, R. Parungao, et al., Burn injury: challenges and advances in burn wound healing, infection, pain and scarring, *Adv. Drug Deliv. Rev.*, **123** (2018), 3–17. https://doi.org/10.1016/j.addr.2017.09.018

4. D. Herndon, F. Zhang, W. Lineaweaver, Metabolic responses to severe burn injury, *Ann. Plast. Surg.*, **88** (2022), 128–131. https://doi.org/10.1097/SAP.0000000000003142

5. A. E. Stoica, C. Chircov, A. M. Grumezescu, Hydrogel dressings for the treatment of burn wounds: An up-to-date overview, *Materials*, **13** (2020), 2853. https://doi.org/10.3390/ma13122853

6. C. Crouzet, J. Q. Nguyen, A. Ponticorvo, N. P. Bernal, A. J. Durkin, B. Choi, Acute discrimination between superficial-partial and deep-partial thickness burns in a preclinical model with laser speckle imaging, *Burns*, **41** (2015), 1058–1063. https://doi.org/10.1016/j.burns.2014.11.018

7. S. Monstrey, H. Hoeksema, J. Verbelen, A. Pirayesh, P. Blondeel, Assessment of burn depth and burn wound healing potential, *Burns*, **34** (2008), 761–769. https://doi.org/10.1016/j.burns.2008.01.009

8. S. Hettiaratchy, R. Papini, Initial management of a major burn: II—assessment and resuscitation, *BMJ*, **329** (2004), 101–103. https://doi.org/10.1136/bmj.329.7457.101

9. F. S. E. Moura, K. Amin, C. Ekwobi, Artificial intelligence in the management and treatment of burns: a systematic review, *Burns Trauma*, **9** (2021). https://doi.org/10.1093/burnst/tkab022

10. H. A. Phelan, J. H. Holmes IV, W. L. Hickerson, C. J. Cockerell, J. W. Shupp, J. E. Carter, Use of 816 consecutive burn wound biopsies to inform a histologic algorithm for burn depth categorization, *J. Burn Care Res.*, **42** (2021), 1162–1167. https://doi.org/10.1093/jbcr/irab158

11. T. Schulz, J. Marotz, S. Seider, S. Langer, S. Leuschner, F. Siemers, Burn depth assessment using hyperspectral imaging in a prospective single center study, *Burns*, **48** (2022), 1112–1119. https://doi.org/10.1016/j.burns.2021.09.010

12. A. G. Monea, K. Baeck, E. Verbeken, I. Verpoest, J. V. Sloten, J. Goffin, et al., The biomechanical behaviour of the bridging vein-superior sagittal sinus complex with implications for the mechanopathology of acute subdural haematoma, *J. Mech. Behav. Biomed. Mater.*, **32** (2014), 155–165. https://doi.org/10.1016/j.jmbbm.2013.12.007

13. M. D. Cirillo, R. Mirdell, F. Sjöberg, T. D. Pham, Improving burn depth assessment for pediatric scalds by ai based on semantic segmentation of polarized light photography images, *Burns*, **47** (2021), 1586–1593. https://doi.org/10.1016/j.burns.2021.01.011

14. N. Brunetti, M. Calabrese, C. Martinoli, A. S. Tagliafico, Artificial intelligence in breast ultrasound: from diagnosis to prognosis-a rapid review, *Diagnostics*, **13** (2022), 58. https://doi.org/10.3390/diagnostics13010058

15. S. A. Suha, T. F. Sanam, A deep convolutional neural network-based approach for detecting burn severity from skin burn images, *Mach. Learn Appl.*, **9** (2022), 100371. https://doi.org/10.1016/j.mlwa.2022.100371

16. C. T. Tchapga, T. A. Mih, A. T. Kouanou, T. F. Fonzin, P. K. Fogang, B. A. Mezatio, et al., Biomedical image classification in a big data architecture using machine learning algorithms, *J. Healthc. Eng.*, **2021** (2021), 1–11. https://doi.org/10.1155/2021/9998819

17. T. S. Hai, L. M. Triet, L. H. Thai, N. T. Thuy, Real time burning image classification using support vector machine, *EAI Endorsed Trans. Context-aware Syst. Appl*, **4** (2017), 4. http://doi.org/10.4108/eai.6-7-2017.152760

18. U. Şevik, E. Karakullukçu, T. Berber, Y. Akbaş, S. Türkyılmaz, Automatic classification of skin burn colour images using texture-based feature extraction, *IET Image Process.*, **13** (2019), 2018–2028. https://doi.org/10.1049/iet-ipr.2018.5899

19. D. P. Yadav, A. Sharma, M. Singh, A. Goyal, Feature extraction based machine learning for human burn diagnosis from burn images, *IEEE J. Transl. Eng. Health. Med.*, **7** (2019), 1–7. https://doi.org/10.1109/JTEHM.2019.2923628

20. S. Lee, H. Ye, D. Chittajallu, U. Kruger, T. Boyko, J. K. Lukan, et al., Real-time burn classification using ultrasound imaging, *Sci. Rep.*, **10** (2020), 1–13. https://doi.org/10.1038/s41598-020-62674-9

21. H. Liu, K. Yue, S. Cheng, W. Li, Z. Fu, A framework for automatic burn image segmentation and burn depth diagnosis using deep learning, *Comput. Math. Methods Med.* , **2021** (2021). https://doi.org/10.1155/2021/5514224

22. J. Karthik, G. S. Nath, A. Veena, Deep learning-based approach for skin burn detection with multi-level classification, in *Advances in Computing and Network Communications: Proceedings of CoCoNet 2020*, **2** (2021), 31–40. https://doi.org/10.1007/978-981-33-6987-0_3

23. C. Pabitha, B. Vanathi, Densemask RCNN: A hybrid model for skin burn image classification and severity grading, *Neural Process Lett.*, **53** (2021), 319–337. https://doi.org/10.1007/s11063-020-10387-5

24. A. Abubakar, H. Ugail, K. M. Smith, A. M. Bukar, A. Elmahmudi, Burns depth assessment using deep learning features, *J. Med. Biol. Eng.*, **40** (2020), 923–933. https://doi.org/10.1007/s40846-020-00574-z

25. C. Jiao, K. Su, W. Xie, Z. Ye, Burn image segmentation based on mask regions with convolutional neural network deep learning framework: more accurate and more convenient, *Burns Trauma.*, **7** (2019). https://doi.org/10.1186/s41038-018-0137-9

26. W. Sun, R. Wang, Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM, *IEEE Geosci. Remote Sens. Lett.*, **15** (2018), 474–478. https://doi.org/10.1109/LGRS.2018.2795531

27. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, **9351** (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

28. X. Liu, Z. Guo, J. Cao, J. Tang, Mdc-net: A new convolutional neural network for nucleus segmentation in histopathology images with distance maps and contour information, *Comput. Biol. Med.*, **135** (2021), 104543. https://doi.org/10.1016/j.compbiomed.2021.104543

29. J. He, Q. Zhu, K. Zhang, P. Yu, J. Tang, An evolvable adversarial network with gradient penalty for covid-19 infection segmentation, *Appl. Soft Comput.*, **113** (2021), 107947. https://doi.org/10.1016/j.asoc.2021.107947

30. N. Mu, H. Wang, Y. Zhang, J. Jiang, J. Tang, Progressive global perception and local polishing network for lung infection segmentation of covid-19 ct images, *Pattern Recognit.*, **120** (2021), 108168. https://doi.org/10.1016/j.patcog.2021.108168

31. C. Zhao, A. Vij, S. Malhotra, J. Tang, H. Tang, D. Pienta, et al., Automatic extraction and stenosis evaluation of coronary arteries in invasive coronary angiograms, *Comput. Biol. Med.*, **136** (2021), 104667. https://doi.org/10.1016/j.compbiomed.2021.104667

32. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

33. K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, et al., Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports, *Acad. Radiol.*, **11** (2004), 178–189. https://doi.org/10.1016/S1076-6332(03)00671-8

AIMS Press