Deep Learning based fully automatic efficient Burn Severity Estimators for better Burn Diagnosis

Joohi Chauhan[1] , Puneet Goyal[1,2]
[1]*Center for Biomedical Engineering,* [2]*Department of Computer Science and Engineering*
Indian Institute of Technology Ropar
Rupnagar, India
{joohi, puneet}@iitrpr.ac.in

Abstract*— Each year, burn injuries lead to several deaths and lifelong disabilities for many others. Timely provided appropriate diagnosis and treatment can reduce sufferings for many, however automated burns diagnosis techniques are still under exploration. Laser Doppler Imaging (LDI) has been found as promising for burns depth assessment, but high costs, delays and portability issues limit its usage in developing automated burns diagnosis methods. The visual images based automated approaches for burn diagnosis have been limitedly explored. This research presents a deep learning based novel approach for burn severity assessment and a new labeled dataset of burn images with varying burn severity that would be made publically available in order to facilitate and advance research for burn severity estimation. As skin characteristics vary across different body regions so will be the burn impact, so we propose customized burn severity estimators (specific to body parts) instead of having a single generic burn severity estimator for the whole human body. Extensive experiments were conducted to evaluate the performance of the proposed approach with different network settings, obtaining competitive results to state-of-the-art methods, despite each customized estimator using a smaller set of images compared to generic one. Also, the experiments suggest that the deep learning based customized estimators perform better than handcrafted features based methods for burns diagnosis.

Keywords*— Burn Injury, Severity Estimation, Classification, Convolutional Neural Network.

I. INTRODUCTION

It is estimated that around 180,000 people die each year and many more suffer prolong physical disabilities because of burn injuries [1]. Automated burn diagnosis can help provide timely treatment that can possibly save many precious human lives and/or alleviate sufferings for many. Also, as the majority of burn injuries occur in low- and middle-income countries [1, 2], where the number of burn units and concerned medical experts is quite low to suitably meet the citizens medical care needs, the automated means for burns diagnosis are very much required to reduce delays and facilitate better treatment.

Burns diagnosis and care require early stage estimation of burn severity (low, moderate or severe) that depends on different factors like cause of burn, burn degree/depth, total body surface area (TBSA %) that is burnt, etc [3,4]. The TBSA % estimation is also needed for evaluating the total fluid requirement of the burn patient. Some research studies suggest laser doppler imaging (LDI) techniques to be quite reliable in assessing burn depth but high costs, delays and portability issues limit its usage in developing automated burns diagnosis methods [5]. As of now, in most places, burn medical professionals and/or paramedical staff have been relying on visual inspection for estimating different burn diagnosis parameters but these traditional subjective methods are found to be slow and not so accurate [5, 6].

Success of deep learning based approaches in many vision and medical images related problems and availability of low cost digital devices having good quality color camera(s) motivate us to investigate and contribute towards color images based automated and efficient burns diagnosis methods using deep learning. Some researchers have earlier contributed towards burn color images segmentation and severity prediction. Also, some research studies had reported that burn depth diagnosis by clinicians using digital images is in good agreement with diagnosis made in person [7, 8] and encouraged remote diagnosis and consultation. A research group from Spain had actively contributed to the automation of burn depth assessment methods from color images [9-11]. They presented a psychophysical experiment using the multidimensional scaling analysis (MDS) and SVM classifier to understand the physical characteristics employed by physicians for burn depth estimation [11]. In their experiment, they used 20 images and used 74 for testing, and the accuracy of 79.73% was achieved where the classes were burns needing grafts and those which did not. Earlier, they have used k-nearest neighbor (kNN) classifier on the same dataset for classifying burn into three depth levels, and 66.2% accuracy was obtained [10]. It was also reported that using their prior approach [9] on the same dataset provided 58.11% accuracy. Their approach, however, required physicians to follow a specific image acquisition protocol for which difficulty in controlling the physicians was noted [10, 11]. Marco et al. had proposed burn wounds images segmentation using tensor decomposition of color images, based on which texture features were extracted for classification [12].

There is very limited research on automated methods for burn severity assessment. Badea et al. [13] proposed an ensemble method, built upon fusing decisions from standard classifiers (SVM and Random Forest) and CNN: ResNet [14], for burn severity estimation. The Histogram of Topographical (HoT) features were extracted and the

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

classifiers such as SVM, RF, LeNet-CNN, and ResNet were trained to differentiate between regular skin patches, light burn patches and serious burn patches [13]. Their proposed system was able to identify light/serious burns with an average precision of 65% using ensemble method. In [13], an extended database of 611 thermal and color images, recorded over several months, was collected, and the color-thermal image pairs had been manually registered; this framework, therefore, is also not easily deployable. Recently, Cirillo et al. [15] presented a burn depth prediction method based on transfer learning. They considered four classes of burns severity and reported an average accuracy of 90.5% using ResNet-101 pre-trained weights based CNN model and a dataset of 458 image (burn) patches derived from the 23 burn images. They used simple transfer learning and it is not evident also if it was ensured that the patches in the training set and those in the test set are not derived from the same individual image for proper evaluation.

The present literature for burn severity estimation from color images is very limited and the performance of automated methods is still an open challenge. Also there are no publicly available benchmark datasets for this challenging burn diagnosis problem. As there are variations in skin anatomy and characteristics on different body regions/parts [16-19] and so will be the differences in the impact of burns on different body parts. Based on this observation, we propose body part specific customized automated models for estimating burn severity. The experimental results show the effectiveness of the proposed approach in predicting the severity level of burn injury from the color images. The main contributions of this paper are as follows:

- We develop a new labeled dataset (Fig. 1) of burn images of four different body parts and having varying burn severity (low, medium and severe). This proposed dataset is used for our experiments and comparative analysis, and it would be made publically available to facilitate further research in this field.
- We propose efficient burn severity estimation (BSE) network models using the predefined deep learning architectures.
- To the best of our knowledge, this is the first work addressing the automated burn severity assessment problem with the customization based on the body parts. As the skin characteristics vary across different body parts, so the body part specific customized BSE models are explored. In this proposed novel framework, the burn images body part classification is first performed, and accordingly, appropriate (body part specific) customized BSE model is selected and applied on the given burn image for predicting the burn severity.
- Qualitative and quantitative experimental results show the remarkable ability of the proposed approach in classifying the body part and its severity from the burn images of patients. Also, comparative analysis of experimental results from state-of-the-art methods on the proposed dataset is presented in this paper.

The remaining paper is organized as follows: Section II discusses the experimental setup including database, data augmentation, proposed methodology, and performance metrics; Experimental results and discussion are given in Section III; and finally, we present the Conclusions of this work in Section IV.

II. Experimental Setup

A. Database

There is no burn images dataset that is publicly available. In general, the availability of medical problems related color images datasets is a known challenge. A *burn-images (BI)* dataset of 432 images is prepared by using semi-automated scripts on the Google search engine, with the average resolution of 754×823 pixels. The poor quality and duplicate images were excluded so as to maintain the integrity. Further, we procured 63 burn images, average resolution of 1254×836 pixels, form iStock [20] for understanding and reporting the effectiveness of the proposed model on an unseen dataset. We name this dataset of 63 test images as *unseen burn images (UBI)* dataset. Furthermore, the initial dataset was labelled manually, based on the visual inspection, for their severity (low, moderate, or severe) in consultation with an experienced senior medical expert (surgeon).

In this work, based on our insight of variation in skin characteristics, impact of burn on different body parts and the susceptibility level of body parts to burn injuries, we consider primarily four different body parts: back, face, hand and inner forearm (IF) [21, 22]. As the final outcome, the dataset is prepared and pre-processed into 4 defined classes of body parts. Fig. 1 presents the sample of burn images of different body parts and Table I shows the overall statistics of the training dataset.

B. Data Augmentation

The availability of more data is always one of the effective and best way to make a neural network generalize better, however the size of dataset is usually limited in practice, especially in medical domain. There are many medical domain and other research studies where the number of available data samples (or say, images) is very limited [23, 24] and the limited dataset is a known challenging aspect for the deep learning based solutions. The number of images in the prepared BI dataset is also limited. However, use of data augmentation (DA) is an effective approach that increases the size of dataset which positively improve the learning potential of network. We enhance the training dataset using label-preserving transformations (rescaling, shear, zoom and horizontal flip) in order to prepare better trained models and as much possibly address the problem of data limitation and overfitting. Data augmentation increases the size of dataset by a factor of 5, hence, raise the number of training samples to 1940. In addition, for BI-dataset before data augmentation, manual data preparation has been done based on image visualization and probability of occurring lesser negative pairs. Based on manual analysis images of each class are blurred, sharpened and mirrored in order to deal with the problem of data limitation in each class and negative pairs.

C. Methodology

We aim to develop efficient automated models for burn

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

severity estimation. We use Convolutional Neural Network (CNN) as the deep CNN have shown very promising results

Fig. 1. Some Sample Images from the proposed databases. The images presented here are of different severity level like 1st image of 2nd row is a low severity burn, 1st row 2nd last image is a moderate burn, and last image of 1st row is a severe burn.

in several complex computer vision and image processing problems and they are increasingly getting adopted for medical image related research challenges. However, it is generally challenging to train a complex convolutional neural network using a small size training set without over-fitting. To deal with such problems, fine-tuning or transfer learning approaches are preferred. In general, fine-tuning modifies the parameter of a pre-trained network and train the model for the new task. We have adopted the similar approach that transfer knowledge between the networks. Fine-tuning [25] an existing network initialized with parameters obtained after training a model on a rich dataset, optimize the parameters with a low learning rate for the new problem to find the local minima. In general, network fine-tuning adapts the shared parameters for making them more discriminative for the new classification problem. However, low learning rate is an approach that indirectly preserve the learned representational structure from the original data used for training network. Research shows that, if the correct set of parameters are used appropriately for fine-tuning network on new dataset, then most of the times, the model outperform the results achieved from a randomly initialized network [26]. We also use known recommended techniques for data regularization: data augmentation, dropout, and early stopping criteria to help address overfitting issues [27-29].

In our work, a large-scale publicly available database, ImageNet, having millions of color images is used in pre-training and developing a CNN that generally consists of sequence of layers: convolutional layers, pooling layer, and activation layers with a ReLU function. The top layer of model is fully connected layer (FC) that helps in performing classification based on the important features learnt and extracted automatically by CNN models. The ResNet50, VGG16 and VGG19 are some existing CNNs well known for their good performance for several image classification problems [14, 30] and are tried in this work as baseline. Our proposed approach for burn severity estimation from color image of a burn patient fine-tunes these pre-trained baseline networks for burn images, and uses a new insight that the body part specific customized estimation models would be likely more efficient as there exists variation in skin properties of different body parts [16-19]. For example, the image characteristics of a low burn on hand may be similar to that of moderate burn face image. The proposed approach for burn severity assessment is a 2-step method - first performing burn images body part classification (BI-BPC), and then accordingly using burn severity estimation (BSE) model. Fig. 2 shows the overview of the proposed body part-specific customized burn severity estimator (C-BSE).

The BI (training) dataset is used for fine tuning the existing pre-trained deep learning networks in order to effectively solve the intended classification problems in this work.

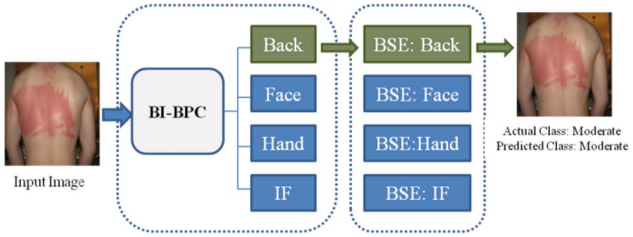

Fig. 2. Overview of the proposed approach – First performing the Burnt image body part classification (BI-BPC), and then using body part specific customized burn severity estimation (BSE) model for predicting the severity.

1) Burn Images – Body Part Classification (BI-BPC): We use the transfer learning based approach for BI-BPC module. Here, we consider three different overlaying CNNs: ResNet50, VGG16, and VGG19, and for each base model, the parameters are first initialized by freezing all other layers except the last set of convolutional layers followed by ReLU and pooling layer. Freezing part of the network supports in preventing overfitting. The trainable layers of the network are trained on BI dataset and utilizing the hyper-parameters obtained from the pre-trained network for extracting the high dimensional features. The last fully connected layer of the

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

network is replaced by a dense layer having 4 nodes, corresponding to 4 different body parts considered in this classification problem.

TABLE I. BI TRAINING DATASET IMAGES IN G-BSE AND C-BSE

Label	No. of Training Images				
	Generic BSE	Customized BSE			
		Back	Face	Hand	I.F.
Severe	132	48	24	36	24
Low	124	48	12	32	32
Moderate	132	52	16	32	32
Total	**388**	**148**	**52**	**100**	**88**

In the learning process, we reserve 30% of the BI dataset images by random selection for the testing purposes and the remaining 70% images are split into training and validation sets followed by data augmentation. Recognizing and considering that the dataset used is not much large, we fine-tune our network with trainable parameter of around *8196* and learning rate for network is set to $lr = 10^{-5}$.

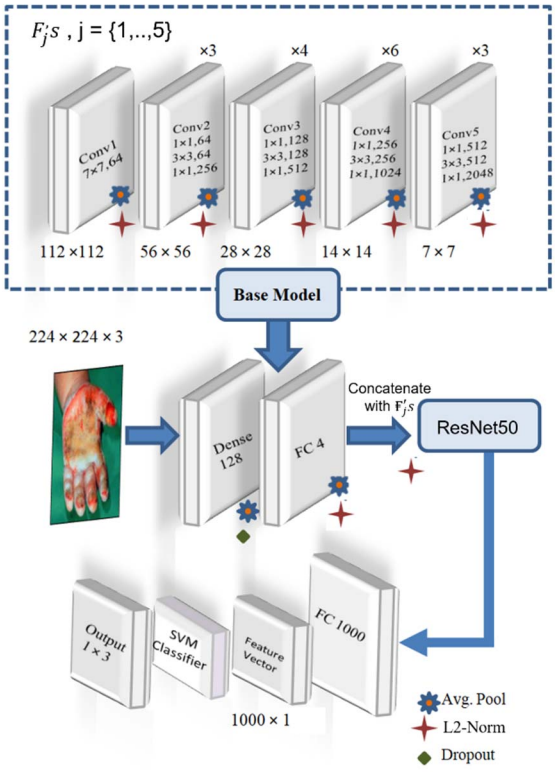

Fig. 3. The architecture of proposed C-BSE Hand model with ResNet50 as base model, consists of (i) burn image body part classification network and (ii) the body part specific burn severity estimation networks.

2) Burn Severity Estimation: For each of the three burn severity labels, the body part specific subsets of BI dataset burn images are small even after augmentation. Using CNN for feature extraction is a viable technique for producing strong learning models; they are capable enough to extract the powerful and complex features form the data for being trained using extensively large image datasets. Considerig all the factors, we utilize pre-trained deep learning models as a feature extractor. For all the network pools, *say n*, in the considered CNN, we extracted the features F_j , *where j= 1 to n* and apply average pooling ($\mathfrak{F}$) on them. The j^{th} feature vector f_j after average pooling is normalized using L_2-Norm, $\mathbb{F}_j$ be the j^{th} feature vector after normalization (1). We also use the features from the last fully connected layer of the network. Fully connected layer features are widely used in many research domain, as it provide strong generalization and semantics-descriptive ability. The extracted features from last fully connected layer F_{FC}, are averaged pooled and normalized using L_2-Norm (2).

$$\text{for } j = \{1,2,\ldots,n\},$$

$$f_j = \mathfrak{F}(F_j),$$

$$\mathbb{F}_j = \| f_j \|_2,$$

$$\text{where } f_j \text{ is the feature vector of } j^{th} \text{ pool.} \quad (1)$$

$$\mathbb{F}_{FC} = \| \mathfrak{F}(F_{FC}) \|_2 \quad (2)$$

Further, we concatenate the feature vectors generated after normalization to get a final feature vector f_s (3) and normalized it further to $\mathbb{F}_s$ (4). The final feature vector from BI training set images is fed into the SVM for severity estimation/classification.

$$f_s = (\mathbb{F}_1 \oplus \mathbb{F}_2 \oplus \mathbb{F}_3 \oplus \ldots \oplus \mathbb{F}_n) \oplus \mathbb{F}_{FC} \quad (3)$$

$$\mathbb{F}_s = \| f_s \|_2 \quad (4)$$

The conventional approach for estimating burns severity based on input color images is to use generic burn severity estimator (G-BSE) that is trained over a (mixed) set of burn images, not specific to a particular body part. However in the proposed customized burn severity estimator (C-BSE) method, we incorporate BSE architecture with BI-BPC such that for every body part class in BI-BPC, there is a specific BSE which has been trained using burn images of that particular body part only. The complete architecture of the proposed C-BSE method is show in Fig. 3. In both, conventional and proposed approaches, the feature extraction and classification method is kept same for the purpose of comparing and presenting the generalizability of the two-step proposed approach.

Here, for body part classification and severity assessment, we have used the cross-entropy loss, denoted as $\mathcal{L}_c$ and $\mathcal{L}_s$, respectively. To train a multi-class single-label classification network, cross-entropy loss is the most popular loss function for training regime; it measures the performance of a classification model where the output is probability between 0 and 1. In case of multiclass classification this loss function is defined as in (6), where M denotes the number of classes, y is the binary indicator and p is the predicted probability.

$$\{\mathcal{L}_c, \mathcal{L}_s\} = -\sum_{c=1}^{M} y_{o,c} \log(p_o, c) \quad (6)$$

Finally, the overall loss $\mathcal{L}_{\text{C-BSE}}$ for the proposed approach 'C-BSE' is the summation of BI-BPC loss $\mathcal{L}_c$ and BSE loss $\mathcal{L}_s$, calculated as:

$$\mathcal{L}_{\text{C-BSE}} = \mathcal{L}_c + \mathcal{L}_s \quad (7)$$

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

The proposed architecture is implemented using the Keras [31] deep learning framework and MATLAB deep learning toolbox [32] on NVIDIA GTX1080 GPU to speed up parameter learning and evaluating the learned model on Intel core 8th Gen i7 CPU. Entire computation and programs of this work run on 64-bit Ubuntu OS with CUDA-10.0 and Tensorflow [33].

D. Performance Evaluation

For the body part classification problem, an input burns image is predicted to be in one of the 4 classes corresponding to 4 different parts considered, furthermore, for the burns severity estimation, there are three classes considered in each body part class. To evaluate the performance of proposed methods, following evaluation metrics are used:

- **Average Accuracy (AA)**: The average of accuracy/ effectiveness of each class of the classifier.
- **Precision$_M$ (P_M)**: Denotes how precise the classifier is, in respect to positive predication. An average of class-wise agreement of the data class labels with those of the classifiers.
- **Recall$_M$ (R_M)**: Measures the ability of a classifier in predicting all the positives. An average per-class effectiveness of a classifier to identify class labels.
- **F1-Score$_M$ ($F1_M$)**: Relations between data's positive labels and those given by a classifier based on a per-class average.

Considering all classes equally, we use macro-averaging while computing these metrics in relation with the elements of confusion matrix.

$AA: \frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$	$P_M: \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$
$R_M: \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$	$F1_M: \frac{2 * P_M * R_M}{P_M + R_M}$

Here tp_i denotes true positive for i^{th} class C_i and l denotes the number of classes in the classification problem considered. The true negatives, false positives and false negatives for C_i are denoted using tn_i, fp_i and fn_i, respectively.

III. Results and Discussion

A. Performance of Burn Images Body Part Classification

As mentioned in Section II, the 70% of BI dataset is used for training and validation, and the rest 30% images of dataset are used as test data. We considered three different cases (case -1, 2 and 3) for unbiased evaluation, similar to 3-fold cross-validation. Three completely disjoint sets, with similar distribution of body part images class-wise (i.e. 4 body part classes), are used as test datasets, and the remaining images followed by augmentation are used for training and validation. Random sampling is performed for the preparation of sets comprising of training, validation and testing subsets, such that for a set, similarity between the within subset classes and cross subsets classes is zero (5), i.e.

$$A_i^j \cap B_i^j \cap C_i^j = \emptyset$$
$$for\ each\ i,j\ where\ i \in \{1,2,3,4\}\ and\ j \in \{0,1,2\}. \quad (5)$$

Similar to 3-fold cross validation, the training in our experiments is performed, keeping the inter-sets testing subset similarity as zero (6), and the cardinality of the same classes of subset of sets is equal (7). The three disjoint sets are described in Table II.

$$\forall\ i \in \{1,2,3,4\},\quad C_i^0 \cap C_i^1 \cap C_i^2 = \emptyset \quad (6)$$

$$\left|A_i^0\right| = |A_i^1| = |A_i^2|,\ \left|B_i^0\right| = |B_i^1| = |B_i^2|,\ \text{and}\ \left|C_i^0\right| = |C_i^1| = |C_i^2| \quad (7)$$

TABLE II. Prepared Sets from BI Dataset for Validating Proposed Methodology

	Subset 1 (Training)	Subset 2 (Validation)	Subset 3 (Testing)
Case 1: Set-0	$A_1^0A_2^0A_3^0A_4^0$	$B_1^0B_2^0B_3^0B_4^0$	$C_1^0C_2^0C_3^0C_4^0$
Case 2: Set-1	$A_1^1A_2^1A_3^1A_4^1$	$B_1^1B_2^1B_3^1B_4^1$	$C_1^1C_2^1C_3^1C_4^1$
Case 3: Set-2	$A_1^2A_2^2A_3^2A_4^2$	$B_1^2B_2^2B_3^2B_4^2$	$C_1^2C_2^2C_3^2C_4^2$

The averaged accuracy results (averaged across three different cases) for each of the four different body part classes, using three different predefined deep learning architectures: ResNet50, VGG16 and VGG19, are presented in Table III. The overall performance of ResNet50 based BI-BPC model is found to be best, and within that the classification performance for inner forearm images (i.e. 92.59%) is observed to be best amongst these four different body parts considered.

TABLE III. Body Part Classification Performance by Proposed BI-BPC Method on BI Test set

Class Labels	**BI-BPC Accuracy** (Mean±Standard Deviation)		
	ResNet50	VGG16	VGG19
Back	85.41 ± 9.55%	74.58 ± 30.83%	79.16 ± 14.43%
Face	83.33 ± 10.67%	78.59 ± 9.62%	72.22 ± 34.70%
Hand	72.73 ± 11.49%	70.94 ± 10.50%	65.76 ± 27.78%
IF	92.59 ± 12.83%	56.74 ± 27.96%	55.56 ± 11.12%

We also evaluate the performance using different handcrafted features and SVM classification methods. The same three different cases were considered while evaluating the efficacy of these feature extraction and machine learning (ML) based popular classification methods. Table IV presents the averaged and standard deviation values of overall accuracy performance for the body part classification obtained by these popular methods and the proposed method on BI test set. Amongst the ML based methods, feature

extraction using HOG is performing better than Haralick and LBP based features, with average overall accuracy of 79.3%. The proposed method is performing significantly better than the traditional machine learning based approaches. As compared to the best performing machine learning approach, i.e. HOG+SVM, an improvement of 4.15% in average accuracy is observed by our method.

TABLE IV. COMPARATIVE ANALYSIS OF OVERALL PERFORMANCE OF BODY PART CLASSIFICATION BY STATE-OF-THE-ART METHODS AND PROPOSED BI-BPC METHOD ON BI TEST SET.

Method	**BI-BPC Overall Accuracy (%)** (Mean±Standard Deviation)
Haralick + SVM	67.46 ± 3.00%
LBP + SVM	76.19 ± 2.38 %
HOG + SVM	79.37 ± 2.48%
Ours	**83.52 ± 8.12%**

B. Performance of Burn Severity Estimation Models

In this section, we discuss the performance of the automated generic burn severity estimator (G-BSE) and the body part specific fully automatic customized burns severity estimator (C-BSE) where the customized BSE, for the given input burn image, is selected in an automated manner using the above discussed ours ResNet50 based proposed BI-BPC method. We also present the performance of semi-automated Truly Customized Burn Severity Estimator (TC-BSE) in which we did not use the automated method for body part classification but *manually inspected* the body part of each of the testing dataset burn images and invoked then *automated* customized BSE corresponding to the body part of that burn image. In this case, body part classification is 100% accurate, as performed manually. Fig. 5 shows the predictions obtained using C-BSE for some of the input test images from BI test set.

In Table V, we report the performance measures (averaged accuracy - AA, precision – P_M, recall – R_M and, F1 measure - $F1_M$) of generic, automatedly customized and truly customized BSEs. It can be observed that for all BSEs, the performance, evaluated using four different metrics, is better for the ResNet50 based BSEs than those using VGG16 or VGG19 for extracting features. In term of average accuracy, the improvement of 3.03%, 9.09%, and 12.12% is noted, respectively, in ResNet50 based G-BSE, TC-BSE, and C-BSE. The performance of generic BSE, where entire training dataset is used without body part classification, is comparatively less than that of customized BSEs. For ResNet50 as feature extractor, the improvement of 10.61/12.2/15.5/13.9% is observed in average accuracy/precision/recall/F1 metric by our proposed method over the generic burn severity estimators.

However, it is also of interest to note that the performance of ResNet50 based C-BSE and TC-BSE is same despite 100% accurate body part classification in TC-BSE, suggesting that some of the images got their severity correctly estimated by customized BSE despite getting misclassified by the body-part classifier, and some of the images were wrongly classified in terms of severity despite using correct body part specific customized burn severity estimator for those images. The last case in Fig. 5 presents a negative case where the inner-forearm image is classified as hand body part images, but the invoked hand body part specific burn severity estimator correctly predicts its severity.

TABLE V. COMPARISON OF AUTOMATED (G,C)-BSE AND SEMI-AUTOMATED TC-BSE IN TERMS OF AVERAGE ACCURACY, PRECISION, RECALL AND F1 MEASURE ON BI TEST SET.

	Method	**ResNet50**	**VGG16**	**VGG19**
AA	G-BSE	0.742	0.712	0.682
	TC-BSE	**0.849**	0.758	0.712
	C-BSE	**0.849**	0.727	0.697
P_M	G-BSE	0.658	0.553	0.524
	TC-BSE	**0.780**	0.637	0.586
	C-BSE	**0.780**	0.587	0.541
R_M	G-BSE	0.621	0.575	0.529
	TC-BSE	**0.776**	0.640	0.570
	C-BSE	**0.776**	0.597	0.546
$F1_M$	G-BSE	0.639	0.563	0.526
	TC-BSE	**0.778**	0.638	0.578
	C-BSE	**0.778**	0.592	0.543

Further, it is encouraging to note that despite the larger number of images for each of the severity category in G-BSE (Table I) compared to any body part specific C-BSEs, the results demonstrate that many of the burn images, especially those of *low* and *moderate* category, are misclassified in generic estimator, but are getting correctly classified by body part-specific C-BSE. An example of such case is shown in Fig. 4, where the actual class of the input image is *moderate*, and correctly classified as *moderate* only by C-BSE, but predicted as of *low severity* burn by G-BSE. It seems that the poor performance of G-BSE is also because of not being able to account well for the shape and skin characteristics variations in different body parts, which is not much a case for the body-part specific C-BSEs.

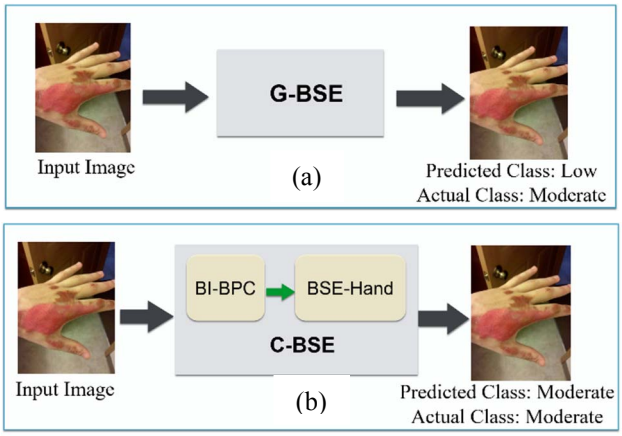

Fig. 4. Prediction on a test image by G-BSE and C-BSE, where the actual class of burn severity is moderate, (a) G-BSE predict the severity of given image as low burn (misclassified), whereas, (b) C-BSE correctly predict the burn severity as moderate.

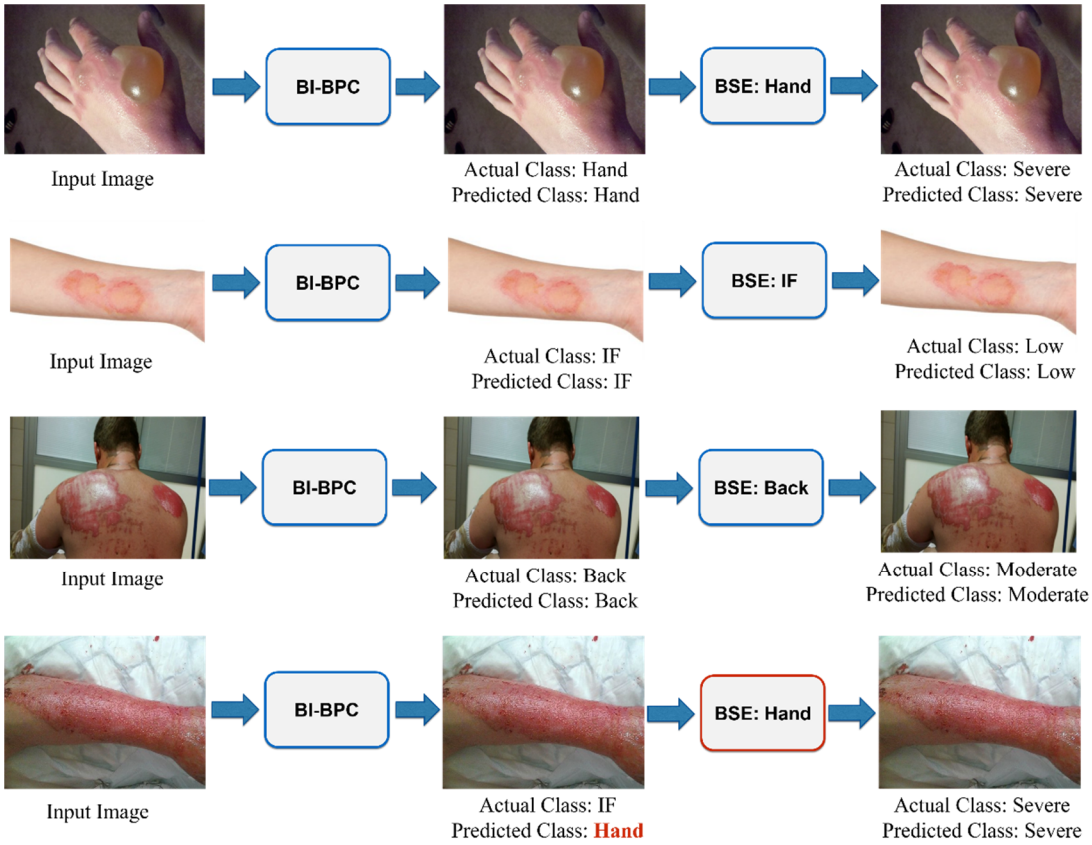

Fig. 5. Qualitative performance of the proposed C-BSE on the sample images of BI test Set

TABLE VI. BURN SEVERITY CLASSIFICATION PERFORMANCE BY PROPOSED METHOD C-BSE ON UBI TEST SET

Metric	ResNet50	VGG16	VGG19
AA	91.53 ± 4.56%	75.66 ± 2.94%	71.43 ± 7.95%
P_M	0.881 ± 0.10	0.622 ± 0.13	0.577 ± 0.06
R_M	0.866 ± 0.07	0.592 ± 0.12	0.535 ± 0.10
$F1_M$	0.873 ± 0.05	0.606 ± 0.11	0.554 ± 0.08

TABLE VII. COMPARATIVE ANALYSIS OF C-BSE PERFORMANCE BY THE STATE-OF-THE-ART METHODS AND PROPOSED METHOD.

Method	C-BSE Overall Accuracy (Mean±Standard Deviation)	
	BI Test sets	**UBI dataset**
Haralick + SVM	60.61 ± 7.06%	60.85 ± 7.24%
LBP + SVM	63.64 ± 7.37%	61.90 ± 4.14%
HOG + SVM	65.15 ± 9.23%	56.61 ± 10.70%
Ours	**84.85 ± 7.36%**	**91.53 ± 4.56%**

For further validating the proposed method, we also tested it on the unseen real world images dataset i.e. UBI dataset. For this UBI dataset also, VGG16 is performing better than \VGG19, but ResNet50 based C-BSE is outperforming VGG16 by 15.53/25.9/27.4/26.7 % in terms of average accuracy/precision/recall/F1 performance metric, as shown in Table VI. The overall average accuracy obtained by the automated proposed C-BSE on UBI dataset is 91.53%. We also perform the comparative analysis between the proposed severity assessment method and the state-of-the-art ML based popular methods. From Table VII, it can be noted that the proposed method is outperforming the handcrafted features based approaches. On BI test set, the average accuracy achieved is 84.85% with an improvement of 19.7%, and on UBI set, it is 91.53% with an improvement of 29.63%, in comparison to the second best performing method.

TABLE VIII. SUMMARY OF DIFFERENT BURN IMAGES CLASSIFICATION METHODS AND THEIR PERFORMANCE ON THEIR DATASETS.

Study	Method	Dataset, # of Images	Accuracy	Precision
Badea *et al.* [13]	Ensemble method (LeNet-CNN, ResNet, HoT+RF)	Color and IR dataset, 611 pairs	60.7%	0.65
Acha *et al.* [9]	Color features based	Color images acquired using specific protocol, 94 images (20+74)	58.11%	-
Acha *et al.* [10]	Psychophysical experiment, MDS analysis and kNN		66.2%	0.73
Serrano *et al.* [11]	Psychophysical experiment, MDS features and SVM		79.73%	0.73
Ours	Body part specific Customised BSEs (using ResNet50 and SVM)	BI dataset, 432	**84.85%**	**0.78**
		(unseen) UBI set, 63	**91.53%**	**0.88**

Table VIII presents the performance of our proposed method and summary of other existing burn image classification methods with their performance on their

datasets. The dataset used for training and testing by all the methods are different, with varying number of images and all these methods, except [11], consider classifying input burn image into one of three types. None of these earlier datasets are available publically. The proposed automated customized BSE method seems to perform reasonably well in comparison, with accuracy on the prepared BI and UBI datasets as 84.85% and 91.53%, respectively.

IV. Conclusion and Future Work

Automated burn severity estimation is helpful for facilitating timely burns diagnosis and treatment. There has been very limited research work in this domain and currently used traditional visual inspection based non-automated approaches are susceptible to subjectivity, errors and time delays. In this work, we present automated and efficient customized models for the burn severity estimation, considering the variations in skin properties across different body parts. We first used a deep learning based network for performing body part classification and then used body part specific customized BSE models. We prepared a labeled dataset of burn images and used it for evaluating our proposed models. ResNet50 pipeline was found to be more efficient and the proposed customized model outperforms the generic BSE models and popular ML based methods. The models can be further enhanced using a larger set of burn images dataset. Further, it would be of interest to develop and investigate the performance of automated customized estimation models for other burns diagnosis parameters (e.g. TBSA %). We hope that the dataset, insights and the results presented in this work would encourage more researchers and the medical experts to pursue further research towards development of better automated burns diagnosis techniques.

Acknowledgment

We thank Dr. Karoon Agrawal, ex-Director and Professor, Department of Burns, Safdarjung Hospital, Delhi, for the useful discussions and assistance in labeling the images.

References

[1] World Health Organisation. "Fact sheets – Burns", Mar 2018 https://www.who.int/news-room/fact-sheets/detail/burns.

[2] C. Mock, M. Peck, M. Peden, and E. Krug, "A WHO plan for burn prevention and care," WHO, Geneva, Switzerland, 2008.

[3] Burns, www.msdmanuals.com/en-in/home/injuries-and-poisoning/burns

[4] Emergency care of moderate and severe thermal burns in adults, https://www.uptodate.com/contents/emergency-care-of-moderate-and-severe-thermal-burns-in-adults.

[5] C. Wearn, K.C. Lee, J. Hardwicke, A. Allouni, A. Bamford, P. Nightingale, and N. Moiemen, "Prospective comparative evaluation study of Laser Doppler Imaging and thermal imaging in the assessment of burn depth," *Burns*, vol. 44, no.1, pp: 124-133, 2018.

[6] A. Karim, K. Shaum and A. Gibson, "Indeterminate-Depth Burn Injury—Exploring the Uncertainty," *Journal Surgical Research*, vol. 245, Jan 2020

[7] L. Roa, T. Gómez-Cía, B. Acha, and C. Serrano, "Digital imaging in remote diagnosis of burns," *Burns*, vol. 25, no. 7, pp. 617–624, 1999.

[8] O. C. Jones, D. I. Wilson, and S. Andrews, "The reliability of digital images when used to assess burn wounds," *J. Telemed. Telecare*, vol. 9, pp. S22–S24, 2003.

[9] B. Acha, C. Serrano, and L. Roa, "Segmentation and classification of burn color images," *EMBC*, vol. 3, pp. 2693-95, 2001.

[10] B. Acha, C. Serrano, I. Fondon, and T. Gomez-Cia, "Burn depth analysis using multidimensional scaling applied to psychophysical experiment data," *IEEE Trans Med Imaging*, vol. 32 (6), pp:1111-20, 2013.

[11] C. Serrano, R. Boloix-Tortosa, T. Gomez-Cia, and B. Acha, "Features identification for automatic burn classification," *Burns*, vol. 41, 2015.

[12] C. D. Marco, M. Robin, S. Folke, and P. Tuan, "Tensor Decomposition for colour Image Segmentation of Burn Wounds," *Scientific Reports Nature*, vol. 9, no. 3291, 2019.

[13] S. M. Badea, C. Vertan, C. Florea, L. Florea, and S. Bădoiu, "Severe Burns Assessment by Joint Color-Thermal Imagery and Ensemble Methods," *IEEE HealthCom, 2016*.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, pp.770-78, Las Vegas, USA, Jun 2016.

[15] M. D. Cirillo, R. Mirdell, F. Sjoberg, and T.D. Pham, "Time-Independent Prediction of Burn Depth Using Deep Convolutional Neural Networks," *Journal of Burn Care and Research*, vol. 40(6), pp. 857-63, 2019.

[16] K. Langer, "On the anatomy and physiology of the skin," *British Journal of Plastic Surgery*, vol. 31, no. 1, pp. 3- 8, 1978.

[17] S. Adabi, M. Hosseinzadeh, S. Noei, S. Conforto, S. Daveluy, A. Clayton, D. Mehregen, and M. Nasiriavanaki, "Universal in vivo textural model for human skin based on optical coherence tomograms," *Scientific Reports* 7, vol.17912, 2017.

[18] T. Igarashi, K. Nishino, and S. K. Nayar, "The Appearance of Human Skin," *Tech. Report:* CUCS-024-05, Columbia University, New York, USA, June 2005.

[19] P. Oltulu, B. Ince, N. Kokbudak, S. Findik, and F. Kilinc, "Measurement of epidermis, dermis, and total skin thicknesses from six different body regions with a new ethical histometric technique," *Turkish Journal of Plastic Surgery*, vol. 26 (2), pp:56-61, 2018.

[20] iStock,https://www.istockphoto.com/in/photos/skin-burn-images.

[21] S. G. Abu-Sittah, E. M. A. Khatib, and A. S. Dibo, "Thermal injury to the hand: review of the literature," *Ann Burns Fire Disasters*, vol. 24, no. 4, pp. 175-185, 2011.

[22] M. Wani, A. M. Ahmad Mir, A.S. Mir, A. Banotra, Y. Watali, and Z. Ahmad, "Epidemiology of burns in teaching hospital of Northern India," *Indian Journal of Burns*, vol. 24, no. 1, pp. 47-52, 2016.

[23] T. Shaikhina, and A. N. Khovanova, "Handling limited datasets with neural networks in medical applications: A small-data approach," Artificial Intelligence in medicine, vol. 75, pp:51-63, 2015.

[24] G. Litjens, T. Kooi, B. K. Bejnordi, A. Setio, F. Ciompi, M. Ghafoorian, J. Laak, B. Ginneken, and C. Sanchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60-88, 2017.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CVPR*, pp. 580-87, June 2014.

[26] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" Advances in Neural Information Processing Systems, pp. 3320–3328, 2014.

[27] J. Kukacka, V. Golkov, and D. Cremers. "Regularization for deep learning: A taxonomy." ˇ arXiv preprint arXiv:1710.10686, 2017.

[28] L. Prechelt. "Automatic early stopping using cross validation: quantifying the criteria," *Neural Networks* 2014; 11(4):761–767.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, 15(1):1929–1958, 2015.

[30] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[31] F. Chollet, Keras, https://github.com/fchollet/keras, 2015.

[32] Mathworks Deep Learning Toolbox: https://in.mathworks.com/help/deeplearning/index.html

[33] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org