Journal of Medical and Biological Engineering (2020) 40:321–333
https://doi.org/10.1007/s40846-020-00520-z

REVIEW ARTICLE

Assessment of Human Skin Burns: A Deep Transfer Learning Approach

Aliyu Abubakar[1,2] · Hassan Ugail[1] · Ali Maina Bukar[1]

Received: 18 November 2019 / Accepted: 14 April 2020 / Published online: 24 April 2020
© The Author(s) 2020

Abstract

Purpose Accurate assessment of burns is increasingly sought due to diagnostic challenges faced with traditional visual assessment methods. While visual assessment is the most established means of evaluating burns globally, specialised dermatologists are not readily available in most locations and assessment is highly subjective. The use of other technical devices such as Laser Doppler Imaging is highly expensive while rate of occurrences is high in low- and middle-income countries. These necessitate the need for robust and cost-effective assessment techniques thereby acting as an affordable alternative to human expertise.

Method In this paper, we present a technique to discriminate skin burns using deep transfer learning. This is due to deficient datasets to train a model from scratch, in which two dense and a classification layers were added to replace the existing top layers of pre-trained ResNet50 model.

Results The proposed study was able to discriminate between burns and healthy skin in both ethnic subjects (Caucasians and Africans). We present an extensive analysis of the effect of using both homogeneous and heterogeneous datasets when training a machine learning algorithm. The findings show that using homogenous dataset during training process produces a biased diagnostic model towards minor racial subjects while using heterogeneous datasets produce a robust diagnostic model. Recognition accuracy of up to 97.1% and 99.3% using African and Caucasian datasets respectively were achieved.

Conclusion We concluded that it is feasible to have a robust diagnostic machine learning model for burns assessment that can be deployed to remote locations faced with access to specialized burns specialists, thereby aiding in decision-making as quick as possible

Keywords Burns · Caucasian skin · African skin · Convolutional neural network · Deep learning · Classification

1 Introduction

Burns are skin injuries caused by heat, radiation and other acute trauma. According to the World Health Organization (WHO), about 300,000 people are dying every year as a result of burn related injuries with a high number of incidences in developing countries [1]. In 2004, nearly 11 million people requiring medical treatment had been reported by WHO. Similarly, the study in [2] finds that nearly half a million Americans are affected by thermal injuries each year with almost 40,000 hospital admissions. The severity of burn injuries varies from superficial burns that heal within 14 days to most complicated deeper (full-thickness) burns that last for more than 3 weeks to heal and requires surgical management [3].

Early patient recovery and healing of burn wounds are of paramount importance that depends on effective and timely assessment. A timely assessment provides an avenue for a decision to be made as early as possible, whether surgery (skin grafting) is required or not [4]. This will ensure shorter hospital stay, reduced expenses and lesser risk of hospital acquired complications. Burns are assessed clinically by observation due to its availability and lesser diagnostic cost. Visual and tactile observation of the wound's descriptions such as sensibility, appearance and capillary blanching is

✉ Aliyu Abubakar
a.abubakar6@bradford.ac.uk

Hassan Ugail
h.ugail@bradford.ac.uk

Ali Maina Bukar
a.m.bukar@student.bradford.ac.uk

[1] Centre for Visual Computing, Faculty of Engineering & Informatics, University of Bradford, Bradford BD7 1DP, UK

[2] Department of Computer Science, Faculty of Science, Gombe State University, Gombe 760214, Nigeria

the common practical approach clinically [5]. The biopsy is an alternative technique used to examine burn depth via the extraction of the sample (biopsied) of the burn wound and examined histologically. Another important non-invasive promising technique for burn assessment is Laser Doppler Imaging (LDI) which measure blood flow and provide a predictive healing potential [6] corresponding to the severity of the burns, thereby providing prompt decision-making.

However, reliability of these approaches suffers diagnostic merit such as inconsistency of assessment by different burn specialist [7], as this is mostly associated with clinical assessment in which its reliability lies on dermatologists' experience. Histological analysis is underpinned by high rates of sampling error and lack of standardized interpretations. LDI's reliability in diagnosing burns in children is affected by movement artifacts, high affordability and operability cost, and requires high level of expertise to operate the device. Therefore, these necessitate the need of an alternative technique that can be robust and effective in terms of accuracy, cost and timely decision-making. In this work, we are proposing the use of a deep transfer learning approach to discriminate whether a given image is burnt or not. To facilitate this, we use datasets from different ethnicities so that a proposed diagnostic approach can be robust and effective.

1.1 Contributions

In this study, we propose the use of deep transfer learning approach to discriminate between human skin burns and healthy skin images in both Caucasian and African patients via the use of transfer learning due to deficient data. The study further provides an extensive analysis of data diversification or inclusion when training a deep learning algorithm. This is crucially important in order to produce a complete robust and less bias diagnostic platform. The rest of the paper is structured as follows: in Sect. 2, we present the highlights of prior work conducted using machine learning approaches and we highlight overview of the relevant convolutional neural networks. In Sect. 3, we present our methodology. In Sect. 4, we present our result and we discuss them. Finally, in Sect. 5, we summarise the findings and discuss possible future directions on which this study can be investigated further.

1.2 Literature Review

Use of machine learning algorithms in solving real world problems is widely applied in different domains [8]. Specifically, deep learning technique has recently achieved remarkable success in different areas such as security [9], traffic forecast [10], agriculture [11] as well as age, gender and face recognition [12–15]. Moreover, in the health sector, deep learning has been applied for image classification and has performed extremely well for detection of diseases [16–19]

A Burns is one of those traumatic injuries subjecting thousands to physical deformities, and in extreme cases, loss of lives affect different body parts such as the face, lower and upper limbs, and neck [20]. The devastating effect is felt severely and causes discomfort to both victims, their families and to the nation as a whole. Recently, a number of researchers have attempted to address burn assessment challenges using machine learning algorithms.

A study by [21] proposed an automation process for the identification and classification of scald burns into different categories (depth) using colour and texture features from LAB images. The experiment was conducted on 50 images in each burn depth category using K-Nearest Neighbour (KNN) and support vector machines (SVM) for the classification. The study shows SVM achieved the highest classification accuracy with 85% in first degree burns as compared to 70% by KNN, 87.5% for SVM in second degree burns as compared to 82% by KNN and 92.5% in third degree burns as compared to 75% by KNN.

Study by [7] used off-the-shelf features extracted by a pre-trained Convolutional Neural Network model and SVM as the classification algorithm for the identification of whether an image contains burns or is healthy. 1360 RGB Caucasian images (equally distributed into burnt and healthy skin) in the proposed study comprising of burn injuries from different body locations or parts. The study achieved a classification accuracy of 99.5% on Caucasian datasets

Another study by the authors in [22] used 74 burn images in LAB colour space from Caucasian patients to discriminate burns using machine learning. The approach used handcrafted features to train a SVM which achieved a classification accuracy of 82.43%.

Additionally, the study reported in [23] proposed a work to classify burns and pressure ulcer wounds in Caucasians. Three different ImageNet pre-trained convolutional neural network models were used as feature extractors while in all the three cases SVM was used as a classifier. The evaluation was carried out via the use of tenfold cross-validation in order to avoid biases which might arise during data splitting process. Interestingly, up to 99% accuracy was recorded via the use of all the features from the ImageNet models.

The study reported in [24] was similarly based on the use of Convolutional Neural Network to predict burn depth in pediatric patients. The study was conducted on 23 burn images based on ground truth defined by experienced clinicians. Original images were then augmented via extraction of a region of interest and resulted to 676 samples out of which 119 are superficial burns, 120 are superficial partial thickness, 108 are intermediate partial thickness, 111 are deep partial thickness, 111 are normal skin images and 107 background images. The authors fine-tuned four different

Convolutional Neural Network models were trained via fine-tuning; ResNet101 yields the maximum performance accuracy of 81.66%, ResNet50 yields 77.79%, GoogleNet yields 73.89%, and VGG-16 achieved 77.53%.

Towards the end, our approach in this study is using state-of-the-art deep learning model (specifically, the pre-trained ImageNet model) to discriminate burns using embedded features representations from diverse ethnicities. To the best of our knowledge, this is the first study that provides extensive experiments and analysis of classifying burns in different ethnic or racial groups.

1.3 Convolutional Neural Network

Convolutional Neural Networks (ConvNet) are machine learning algorithms inspired by the human brain and are used as architecture for classification such as image recognition. ConvNet architecture generally consists of a series of different layers, where in each layer 2D array of pixels (feature maps) are produced which serves an input to the next layer. Training of ConvNet architecture was a bottleneck to researchers due to inability to access a huge amount of data and powerful computational machines until around 2010 when a large repository of images called ImageNet [25] was made available. The ConvNet architecture is fundamentally made up of the following layers [26]:

- *Input layer* this is where data are passed into the network. The data can be a raw image pixel or their transformations.
- *Convolutional layer* this layer contains fixed size filters arranged in series performing convolution operations and producing what is referred to as a feature map.
- *Pooling layer* this is where the dimensions of the feature map produced by convolutional layer are reduced, thereby allowing the network to focused on the most important features.
- *Rectified Linear Unit (ReLU)* this layer is responsible for removing all negative values by applying a non-linear function to the output of the previous layer and setting them to zero.
- *Fully connected Layer* this is the layer where the high-level reasoning of the patterns generated by the previous layers is done. All the activations in the previous layer have full connection to neurons in this layer. For feature extraction using pre-trained ConvNet model, features are generated here and used to train another classification algorithm
- *Loss layer* this is where the deviation between the true and the predicted labels is penalized. This is normally the last layer of the ConvNet, and various loss functions are used depending on the task. Example of loss functions includes SoftMax, Cross-Entropy and Sigmoid

Deep ConvNet models become poplar and the research domain receive more recognition due data availability and the computational resources from 2010. Among the most common ConvNet models from 2010 to date includes AlexNet, VGG-16 and ResNet-50. AlexNet model is composed of 5 convolutional layers with an interweaved max-pooling layers and 3 fully-connected layers proposed by the authors of [27] from the University of Toronto in Canada. The firs convolutional layer was configured to take an input size 224×224 and equipped with 96 filters of size 11×11 with a stride of 4. The output of the first layers goes into the second layer as input (after the pooling operation) equipped with 256 filters of size 5×5. The third, fourth and fifth layers contain 384, 384 and 256 filters respectively with the same filter size of 3×3 while the fully connected layers have 4096 neurons each and the soft-max layer as the final layer for classification.

In 2014, Visual Geometry Group (VGG) from Oxford University made another breakthrough in image classification task using a sixteen layered ConvNet model; thirteen convolutional and three fully connected layers [28]. Apart from been deeper than AlexNet, the size of receptive fields was significantly reduced in both convolutional and pooling layers to 3×3 and 2×2 respectively and stride of 2 was maintained throughout. Similar to AlexNet model, the soft-max layer was used as the final classification layer.

GoogleNet achieved a remarkable performance in 2014 [29]. It has a total of 22 layers or 27 layers (pooling layers inclusive). The architecture of the model comprises of parallel layers of convolution with different filters, the output of such parallel convolutional layers is then concatenated as input to subsequent layer. Unlike previously proposed models, GoogleNet is also equipped with 1×1 convolution for dimensionality reduction. GoogleNet outperformed all other proposed models in the ILSVRC competition in 2014.

However, deeper ConvNet encountered several difficulties during training which includes vanishing gradients and accuracy degradation, which was one of the issues faced by VGG. When the network is deep, the gradient shrinks to zero from where the loss function was computed, thereby resulting in a network not learning anything. As such, among the researchers who proposed a pipeline to address the challenges by allowing the network to go deeper with the increase in performance includes Residual Network from Microsoft [30], and won the ILSVRC competition in 2015. Residual Network (known as ResNet) uses skip connections to allows a copy of the gradient to be passed to the subsequent layer without passing through other weight layers as depicted in Fig. 1.

Towards the end, our proposal in this paper is straightforward, which involves automatic recognition of burns using deep learning.

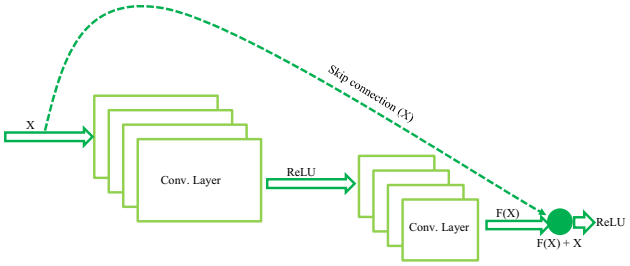

Fig. 1 Illustration of the residual block

2 Materials and Methodology

ConvNet is one of the most known machines learning algorithms these days, and their use has been exploited in medical image analysis. They are supervised machine learning techniques that can extensively extract discriminatory image features using a considerable amount of training data. Basically, deploying ConvNet is carried out using three approaches; training from a ConvNet from scratch, fine-tuning an existing ConvNet and off the shelf features. Training from scratch requires powerful computational machines and huge data, which is a very challenging task. Fine-tuning and off the shelf features approaches are referred to as transfer learning. Fine-tuning involves customizing top-most layers of an existing learned ConvNet and freezing lower layers, in which features extracted from the frozen layers are used to trained top-most customized or added layers using the new dataset. Lastly, classification algorithms such as Support Vector Machines (SVM) and Decision Trees (DT) can be trained using off the shelf features extracted by freezing learned layers of the existing ConvNet.

Generally, there are several existing pre-trained models that are used for transfer learning, for example, VGG16, VGG19, ResNet50, ResNet101.

For the problem domain, we propose in this study, and we utilised ResNet50 pre-trained model. Basically, our methodology here uses a fine-tuning approach in which initial layers of ResNet50 were frozen to extract useful features, and subsequently, top-most substituted layers were trained using those features from the initial layers. In the next subsections, we presented the datasets, fine-tuning and experimental frameworks.

2.1 Data Collection

Datasets were collected from patients of different ethnicities involving Caucasians and Africans. The Caucasian datasets were collected in Bradford hospital, the United Kingdom and the African datasets were collected from Federal Teaching Hospital Gombe, Nigeria. 1360 Caucasian images were successfully collected, which contained 680 burn images and 680 healthy skin images. African dataset contains 270 burn images and 270 healthy skin images, totaling 540 images. Figure 2 shows samples of burns from the two ethnicities. The database images are composed of both pediatric and adult patients as well as different burn complexities.

2.2 Pre-processing

Prior to the development of this research, all datasets from both Caucasians and Africans contain regions that are not relevant to burns identification. In order to diminish the risk of distorting final results, such as the regions were carefully cropped out, and the images were normalized to enhance homogeneity.

2.3 Fine-Tuning Scenario

Pre-trained ConvNet models become very useful due to the ability to transfer their internal deep representations in solving various recognition tasks in another domain faced with challenges of enormous data availability. They can be used as feature extractors or modify (fine-tune) towards a new task. Generic features of the images such as edges and blobs are captured from early layers of a pre-trained model while the later layers get more specific image details. Fine-tuning a pre-trained ConvNet model is to copy all the layers of the model excluding the last layer and replace it with a new specific layer that corresponds to the number of classes in the new domain. Figure 3 below depicts an illustration of the scenario.

We freeze the lower layers of the ResNet50 as depicted in Fig. 3 while customizing the top layers. The customized ResNet50 has a fully connected layer with 512 nodes which passes through a Rectified Linear activation layer (ReLU), then a dropout layer was added to ensure good generalization. Figure 4 shows a fragment of code used in modifying the last layer of the pre-trained ConvNet model. The original model outputs 1000 different object categories while the new customized model output 2 classes (i.e. healthy skin or unhealthy skin).

2.4 Training Setting

We artificially enlarge the training dataset via a process called augmentation. The augmentation strategy includes random resize-crop, random aspect ratio, random rotation, horizontal flip, and center crop. This is in addition to avoid training on a very small database which may lead to overfitting. 80% of the dataset was allocated for training and the remaining 20% for validation.

We utilised a free-access jupyter environment called Google collaborator (Colab) which allows development and training deep learning algorithms using on-demand

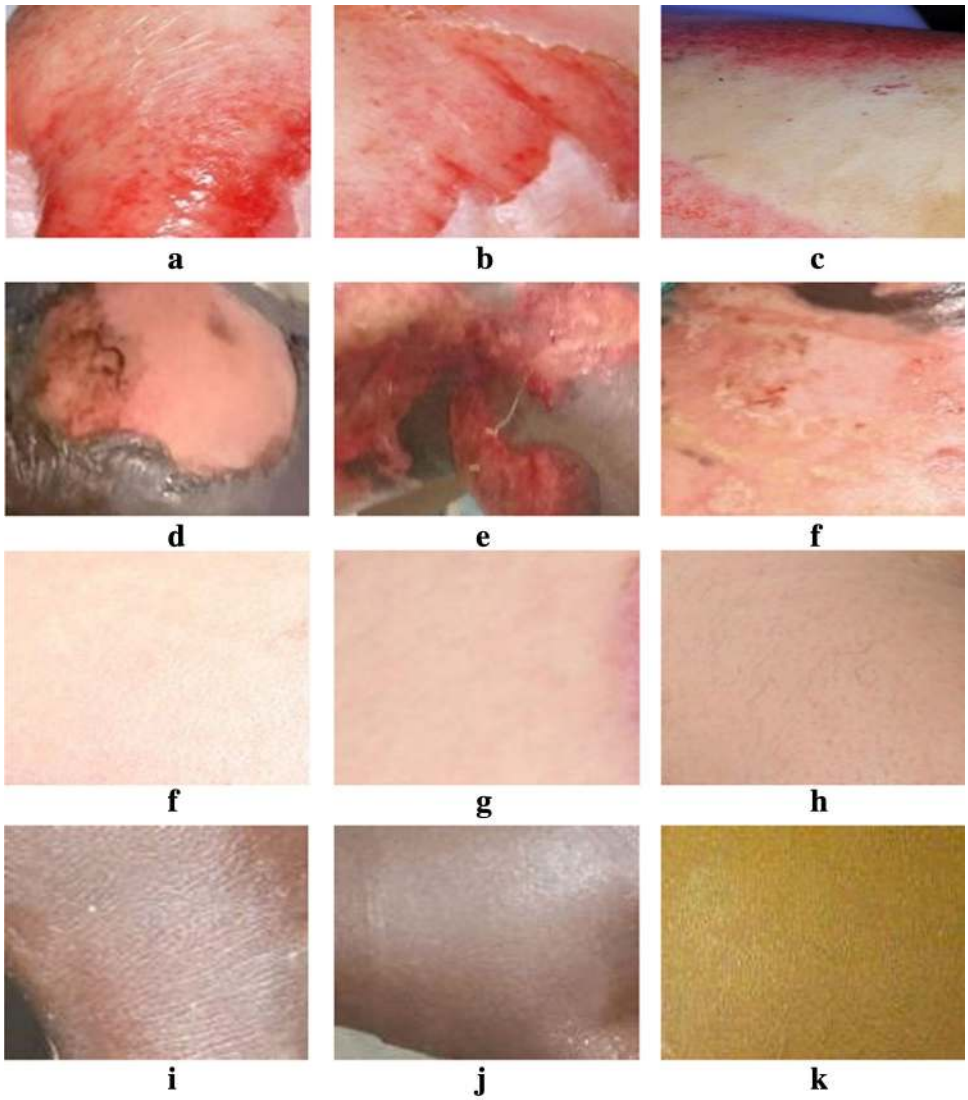

Fig. 2 Dataset samples: **a**–**c** are burn samples from Caucasian patients; **d**–**f** are burn samples from African patients; while **f**–**k** are healthy skin samples from Caucasian and African patients respectively

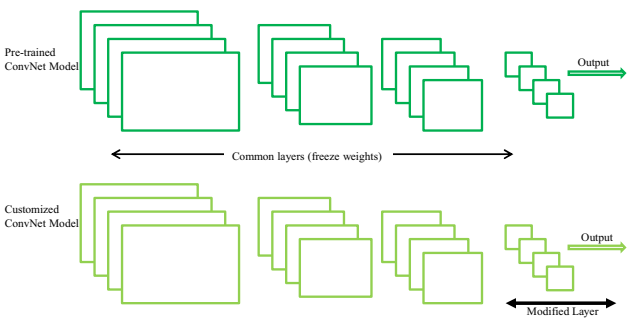

Fig. 3 Illustration of fine-tuning scenario

powerful computing resources. Colab gives free access to python libraries and to develop deep learning applications to be run on hardware equipped with NVIDIA Tesla K80, 12 GB RAM 2496 CUDA cores @ 560 MHz.

```
base_model=ResNet50()
model=base_model.output
model=GlobalAveragePooling2D()(model)
model=Dense(512,activation='relu')(model)
model=Dropout(0.6)(model)
model=Dense(512,activation='relu')(model)
model=Dense(classes, activation='sigmoid')(model)
```

Fig. 4 Fragment of code of the fine-tuned layer

3 Results and Discussion

Figure 5 presents the experimental framework in which features extracted by the frozen layers of ResNet50 were used in training the newly added two dense layers and a classification layer.

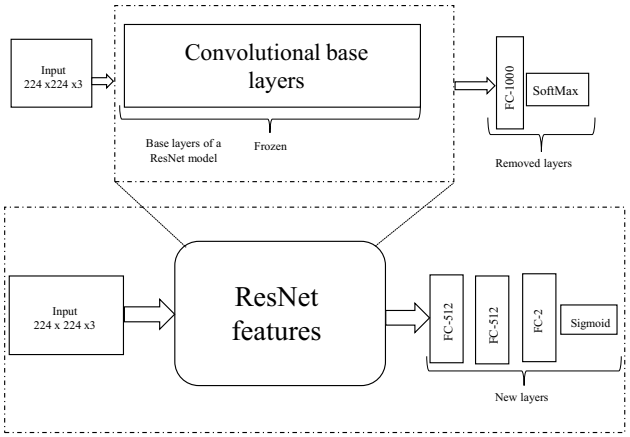

Fig. 5 Illustration of our experimental framework

```
Epoch: 174/200
Epoch : 173, Training: Loss: 0.0380, Accuracy: 98.6213%,
              Validation : Loss : 0.0153, Accuracy: 99.2647%, Time: 12.5258s
Epoch: 175/200
Epoch : 174, Training: Loss: 0.0212, Accuracy: 99.2647%,
              Validation : Loss : 0.0351, Accuracy: 98.8971%, Time: 12.4921s
Epoch: 176/200
Epoch : 175, Training: Loss: 0.0213, Accuracy: 99.1728%,
              Validation : Loss : 0.0237, Accuracy: 99.2647%, Time: 12.5278s
Epoch: 177/200
Epoch : 176, Training: Loss: 0.0131, Accuracy: 99.4485%,
              Validation : Loss : 0.0195, Accuracy: 99.2647%, Time: 12.4817s
Epoch: 178/200
Epoch : 177, Training: Loss: 0.0073, Accuracy: 99.7243%,
              Validation : Loss : 0.0114, Accuracy: 99.2647%, Time: 12.5302s
Epoch: 179/200
Epoch : 178, Training: Loss: 0.0345, Accuracy: 98.9890%,
              Validation : Loss : 0.0398, Accuracy: 98.8971%, Time: 12.5065s
Epoch: 180/200
Epoch : 179, Training: Loss: 0.0191, Accuracy: 99.6324%,
              Validation : Loss : 0.0193, Accuracy: 99.2647%, Time: 12.5269s
Epoch: 181/200
Epoch : 180, Training: Loss: 0.0249, Accuracy: 99.1728%,
              Validation : Loss : 0.0248, Accuracy: 98.8971%, Time: 12.4675s
Epoch: 182/200
Epoch : 181, Training: Loss: 0.0208, Accuracy: 99.1728%,
              Validation : Loss : 0.0124, Accuracy: 99.6324%, Time: 12.4645s
Epoch: 183/200
Epoch : 182, Training: Loss: 0.0113, Accuracy: 99.7243%,
              Validation : Loss : 0.0207, Accuracy: 99.6324%, Time: 12.5750s
```

Fig. 6 Training process on the Caucasian dataset

We conducted a series of experiments using Caucasian, African and combination of both (global) datasets to reveal the best training approach using diverse feature representations that to have a robust diagnostic model. Figure 6 shows the training process in which only datasets from Caucasian patients are used in training the algorithm.

Training and validation losses are depicted in Fig. 7. It is obvious that there is no or minimal over-fitting challenges, the trained model has fitted well with the training dataset, and the performance on the validation set is good. Both training and validation losses settled rapidly from early epochs, and the impressive performance was maintained up to the last epoch.

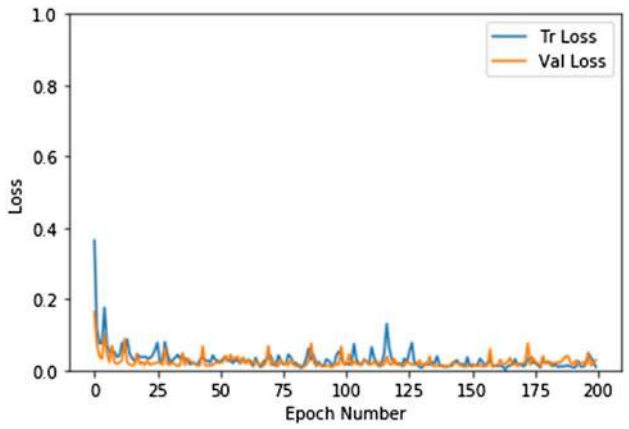

Fig. 7 Training and validation loss using the Caucasian dataset

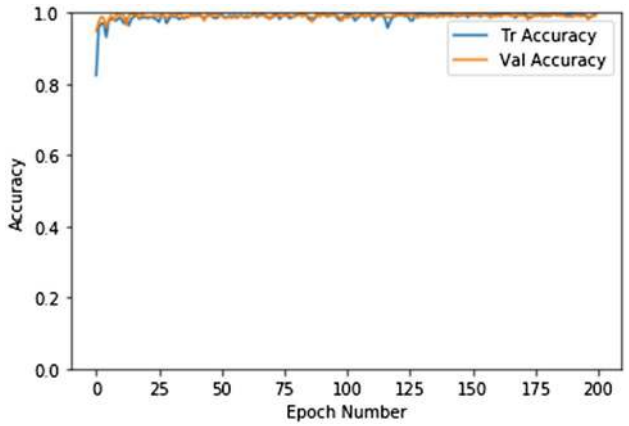

Fig. 8 Training and validation accuracy on the Caucasian dataset

Figure 8 depicts the training and validation accuracies, maximum validation accuracy of 99.6% was achieved, the performance stabilized even before epoch 200.

On the other side, Fig. 9 shows the training process using the dataset from African patients. The maximum validation accuracy achieved by the trained model is 96.4%.

Figure 10 shows training and validation losses, while Fig. 11 shows training and validation accuracy using datasets from African patients. We recorded an impressive result using the proposed approach with slight poor generalization compared to the previous result using Caucasian images. We attributed the high misclassification rate to poor resolutions of some of the images and poor or uncontrolled illumination during the acquisition process. This raises an alarm that the trained machine learning algorithm on Caucasian datasets can be biased when African skin is tested. This is one of the major limitations of previous studies. We believe the research is not satisfactory without taking ethnic or racial representations during the training process. More than

```
Epoch: 180/200
Epoch : 179, Training: Loss: 0.0704, Accuracy: 97.1429%,
                Validation : Loss : 0.2769, Accuracy: 95.0000%, Time: 4.7014s
Epoch: 181/200
Epoch : 180, Training: Loss: 0.0429, Accuracy: 98.5714%,
                Validation : Loss : 0.2508, Accuracy: 95.0000%, Time: 4.7083s
Epoch: 182/200
Epoch : 181, Training: Loss: 0.0642, Accuracy: 97.3214%,
                Validation : Loss : 0.2313, Accuracy: 95.0000%, Time: 4.7318s
Epoch: 183/200
Epoch : 182, Training: Loss: 0.0407, Accuracy: 99.2857%,
                Validation : Loss : 0.2377, Accuracy: 95.0000%, Time: 4.7359s
Epoch: 184/200
Epoch : 183, Training: Loss: 0.0557, Accuracy: 98.2143%,
                Validation : Loss : 0.3365, Accuracy: 93.5714%, Time: 4.7342s
Epoch: 185/200
Epoch : 184, Training: Loss: 0.0361, Accuracy: 98.5714%,
                Validation : Loss : 0.2082, Accuracy: 95.0000%, Time: 4.7199s
Epoch: 186/200
Epoch : 185, Training: Loss: 0.0501, Accuracy: 98.0357%,
                Validation : Loss : 0.3300, Accuracy: 95.0000%, Time: 4.7088s
Epoch: 187/200
Epoch : 186, Training: Loss: 0.0658, Accuracy: 98.0357%,
                Validation : Loss : 0.2893, Accuracy: 95.0000%, Time: 4.7513s
Epoch: 188/200
Epoch : 187, Training: Loss: 0.0495, Accuracy: 97.8571%,
                Validation : Loss : 0.3691, Accuracy: 92.8571%, Time: 4.7150s
Epoch: 189/200
Epoch : 188, Training: Loss: 0.0308, Accuracy: 98.9286%,
                Validation : Loss : 0.1859, Accuracy: 96.4286%, Time: 4.7354s
```

Fig. 9 Training process using the African dataset

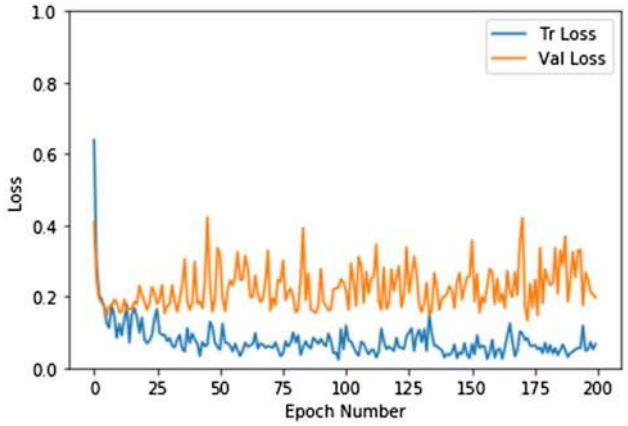

Fig. 10 Training and validation loss on African dataset

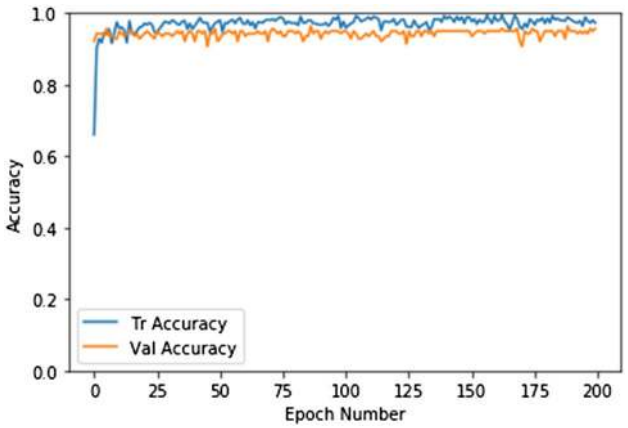

Fig. 11 Training and validation accuracy using the African dataset

```
Epoch 168/200
37/36 [==============================] - 6s 155ms/step - loss: 0.0139 - acc: 0.9922 - val_loss: 0.8406 - val_acc: 0.8214
Epoch 169/200
37/36 [==============================] - 6s 155ms/step - loss: 0.0032 - acc: 0.9991 - val_loss: 0.6770 - val_acc: 0.8512
Epoch 170/200
37/36 [==============================] - 6s 155ms/step - loss: 0.0037 - acc: 0.9982 - val_loss: 0.8972 - val_acc: 0.8244
Epoch 171/200
37/36 [==============================] - 6s 155ms/step - loss: 0.0025 - acc: 0.9991 - val_loss: 0.5133 - val_acc: 0.8750
Epoch 172/200
37/36 [==============================] - 6s 155ms/step - loss: 0.0023 - acc: 0.9982 - val_loss: 0.9124 - val_acc: 0.8363
```

Fig. 12 Showing training process using Caucasian dataset and validated using the African dataset

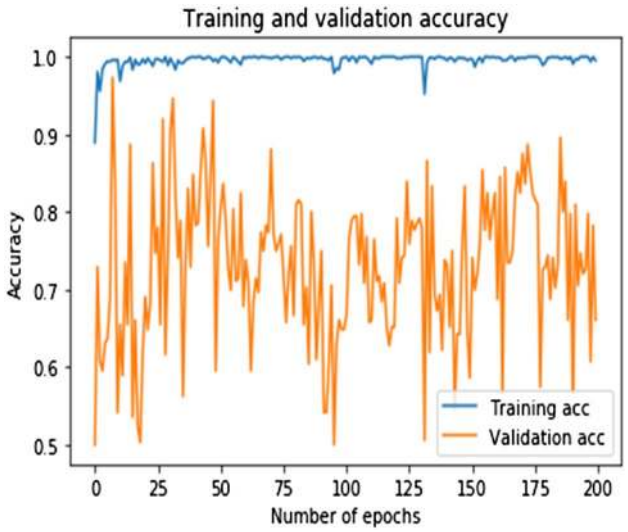

Fig. 13 Training and validation accuracy of a model trained on the Caucasian dataset and validated on the African dataset

90% of the epidemiology of burn related injuries occur in low/middle-income countries such as those in African and Asian counties [31]. Therefore, we further explore whether a trained model on a specific ethnic dataset can provide good identification accuracy on other racial datasets.

Lack of racial or diverse ethnic inclusion during the training process of a machine learning algorithm tends to produce the unrealistic model as shown in Fig. 12 in which training process using Caucasian datasets is depicted, and Fig. 13 shows training and validation accuracy where Fig. 14 shows the corresponding training and validation loss. The validation accuracy is clearly not impressive using data from African patients. The model seems to have overfit. A similar finding is obtained when the model is trained using images from African patients, the validation accuracy using Caucasian data shows poor generalization in which the model tends to be biased as depicted in Figs. 15, 16 and 17.

The results presented in Figs. 13 and 16 clearly indicates that ConvNets are biased in such a way that they recognise a particular racial data they were trained with. This phenomena of recognizing a racial group while failing to recognise another racial group is termed as 'other-race effect' [32]. Taking Figs. 24 and 27 into consideration, this tells us that racial composition in the training datasets tends to produce

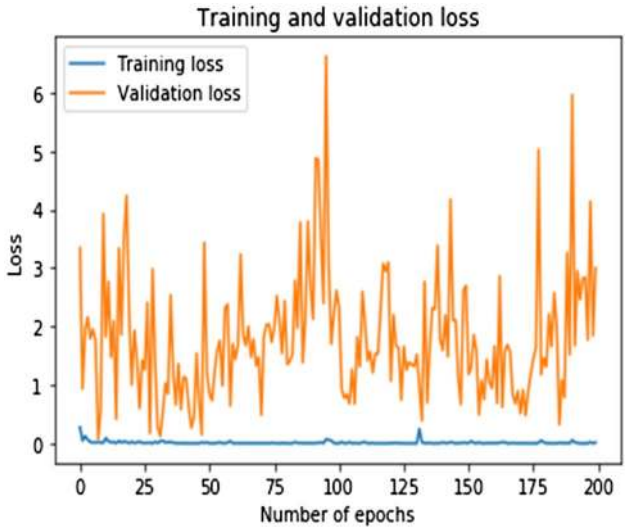

Fig. 14 Training and validation loss of a model trained on the Caucasian dataset and validated using African dataset

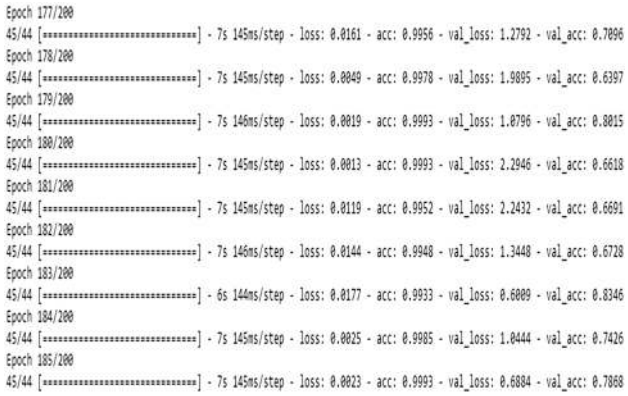

Fig. 15 Showing training process using African dataset and validated using the Caucasian dataset

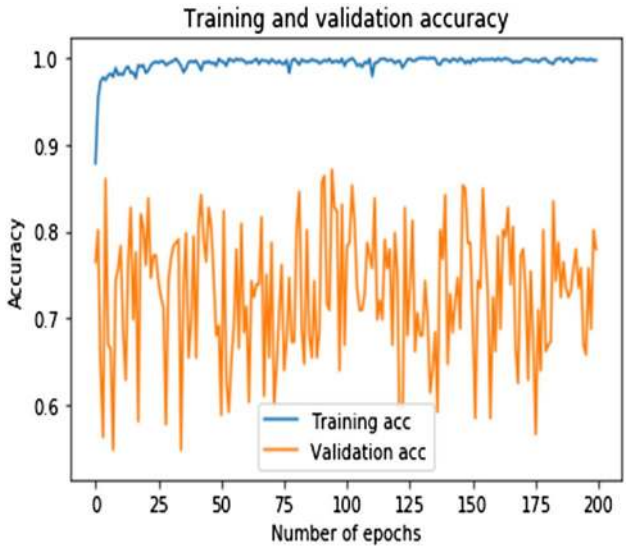

Fig. 16 Training and validation accuracy of a model trained on African dataset and validated using the Caucasian dataset

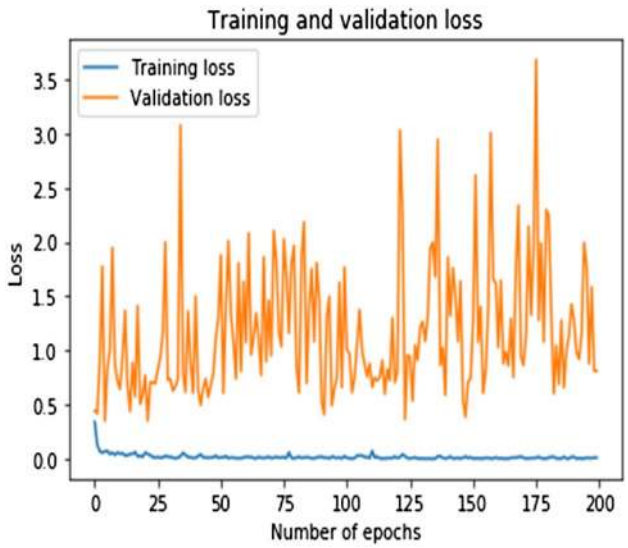

Fig. 17 Showing training and validation loss using African dataset and validated using the Caucasian dataset

a robust system that can be deployed and be utilized effectively while diminishing bias.

We have seen so far how good ConvNets are in discriminating burns in both Caucasian and African skin types. We have also seen how poor the performance of a ConvNet trained using just a specific racial data is. At this point, both datasets from Caucasians and Africans were put together, forming a new type of dataset, which we simply designated as the global dataset. The following depicted Figs. 18, 19 and 20 show the results of a trained model using global datasets, in which validation dataset contains another explicit global data. The result is quite impressive with the validation accuracy of up to 98%.

We further introduce reshuffling operation during the training process using global datasets in order to diminish variance and to ensure the model remain robust and overfit less. Figures 21, 22 and 23 show the training process—with the training and validation losses, and training and validation accuracies respectively. The performance reaches up to 99% (see Fig. 23) outperforming the previous outcome depicted in Fig. 20.

Depicted Figs. 24, 25 and 26 show result in which global dataset is used during the training while validation datasets contain only Caucasian dataset. This shows hard-encoded data representation or embedded features from diverse ethnicities provides effective identifications of burns regardless

```
Epoch: 180/200
Epoch : 179, Training: Loss: 0.0704, Accuracy: 97.1429%,
                Validation : Loss : 0.2769, Accuracy: 95.0000%, Time: 4.7014s
Epoch: 181/200
Epoch : 180, Training: Loss: 0.0429, Accuracy: 98.5714%,
                Validation : Loss : 0.2508, Accuracy: 95.0000%, Time: 4.7083s
Epoch: 182/200
Epoch : 181, Training: Loss: 0.0642, Accuracy: 97.3214%,
                Validation : Loss : 0.2313, Accuracy: 95.0000%, Time: 4.7318s
Epoch: 183/200
Epoch : 182, Training: Loss: 0.0407, Accuracy: 99.2857%,
                Validation : Loss : 0.2377, Accuracy: 95.0000%, Time: 4.7359s
Epoch: 184/200
Epoch : 183, Training: Loss: 0.0557, Accuracy: 98.2143%,
                Validation : Loss : 0.3365, Accuracy: 93.5714%, Time: 4.7342s
Epoch: 185/200
Epoch : 184, Training: Loss: 0.0361, Accuracy: 98.5714%,
                Validation : Loss : 0.2082, Accuracy: 95.0000%, Time: 4.7199s
Epoch: 186/200
Epoch : 185, Training: Loss: 0.0501, Accuracy: 98.0357%,
                Validation : Loss : 0.3300, Accuracy: 95.0000%, Time: 4.7088s
Epoch: 187/200
Epoch : 186, Training: Loss: 0.0658, Accuracy: 98.0357%,
                Validation : Loss : 0.2893, Accuracy: 95.0000%, Time: 4.7513s
Epoch: 188/200
Epoch : 187, Training: Loss: 0.0495, Accuracy: 97.8571%,
                Validation : Loss : 0.3691, Accuracy: 92.8571%, Time: 4.7150s
Epoch: 189/200
Epoch : 188, Training: Loss: 0.0308, Accuracy: 98.9286%,
                Validation : Loss : 0.1859, Accuracy: 96.4286%, Time: 4.7354s
```

Fig. 18 Training process on the global dataset

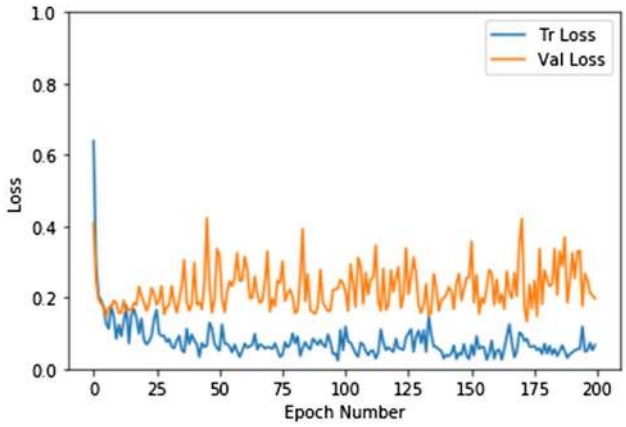

Fig. 19 Training loss using global datset and validation loss using the global dataset

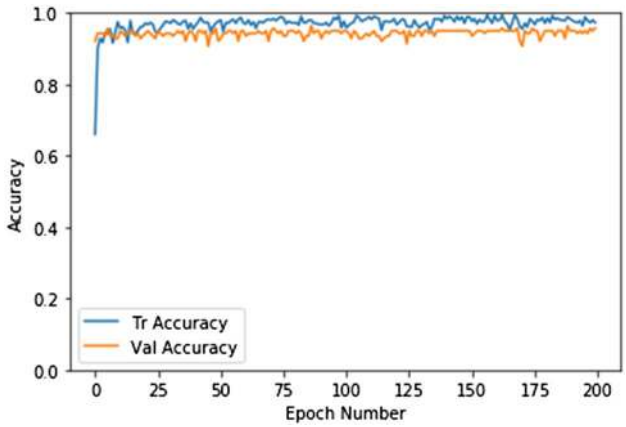

Fig. 20 Training accuracy using global dataset and validation accuracy using the global dataset

```
Epoch : 126, Training: Loss: 0.0426, Accuracy: 98.6650%,
                Validation : Loss : 0.0370, Accuracy: 98.3010%, Time: 27.3099s
Epoch: 128/200
Epoch : 127, Training: Loss: 0.0548, Accuracy: 98.3617%,
                Validation : Loss : 0.0379, Accuracy: 98.3010%, Time: 27.9875s
Epoch: 129/200
Epoch : 128, Training: Loss: 0.0454, Accuracy: 98.4830%,
                Validation : Loss : 0.0366, Accuracy: 98.3010%, Time: 27.7597s
Epoch: 130/200
Epoch : 129, Training: Loss: 0.0457, Accuracy: 98.2403%,
                Validation : Loss : 0.0372, Accuracy: 98.7864%, Time: 27.7963s
Epoch: 131/200
Epoch : 130, Training: Loss: 0.0287, Accuracy: 99.0291%,
                Validation : Loss : 0.0491, Accuracy: 97.8155%, Time: 27.3459s
Epoch: 132/200
Epoch : 131, Training: Loss: 0.0323, Accuracy: 99.0291%,
                Validation : Loss : 0.0381, Accuracy: 98.3010%, Time: 27.5934s
Epoch: 133/200
Epoch : 132, Training: Loss: 0.0508, Accuracy: 97.9369%,
                Validation : Loss : 0.0376, Accuracy: 98.3010%, Time: 27.8206s
Epoch: 134/200
Epoch : 133, Training: Loss: 0.0298, Accuracy: 98.7257%,
                Validation : Loss : 0.0229, Accuracy: 99.2718%, Time: 27.9802s
Epoch: 135/200
Epoch : 134, Training: Loss: 0.0476, Accuracy: 98.2403%,
                Validation : Loss : 0.0216, Accuracy: 99.5146%, Time: 27.2935s
```

Fig. 21 Training process using the reshuffled global dataset

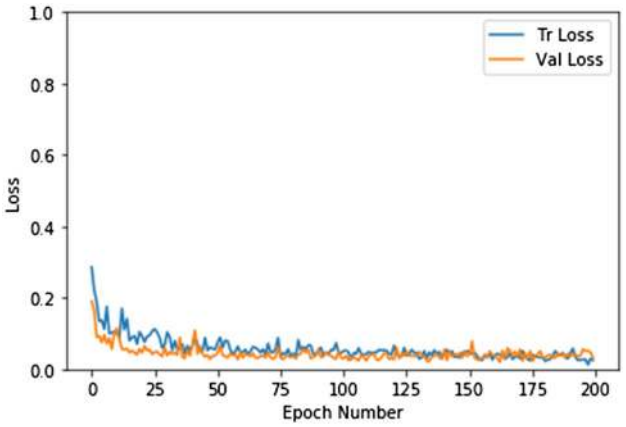

Fig. 22 Training and validation loss using the reshuffled global dataset

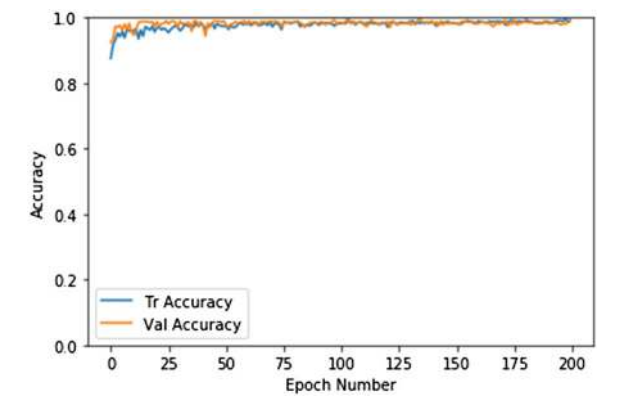

Fig. 23 Training and validation accuracy using the reshuffled global dataset

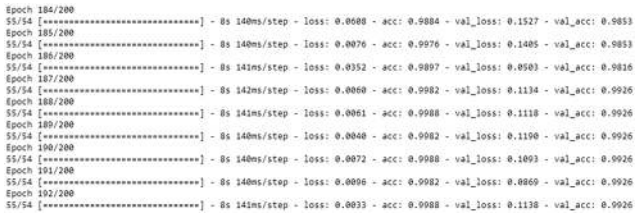

Fig. 24 Showing training process using global datasets and validation using the Caucasian datasets

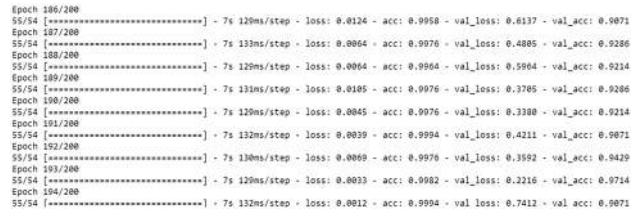

Fig. 27 Showing training process using global datasets and validation using the African datasets

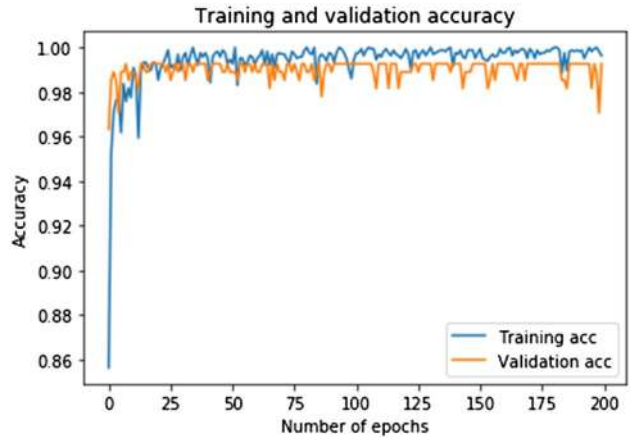

Fig. 25 Showing training accuracy using global datasets and validation accuracy using the Caucasian datasets

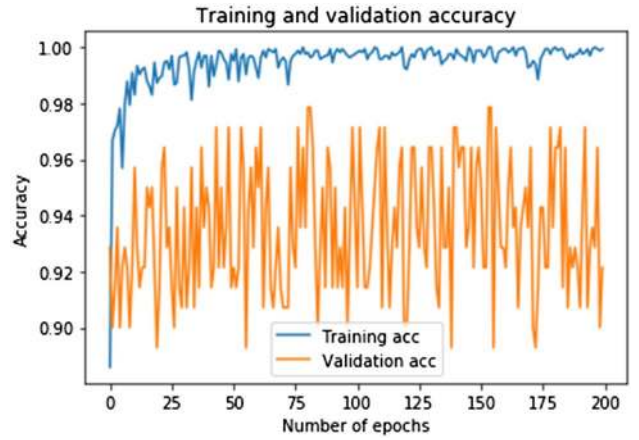

Fig. 28 Showing training accuracy using global datasets and validation accuracy using the African datasets

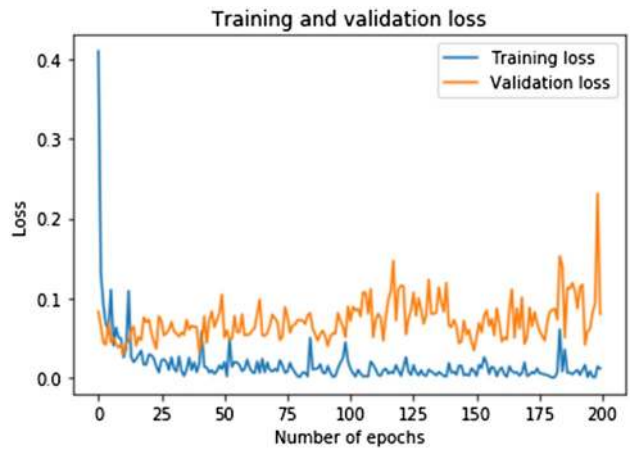

Fig. 26 Showing training loss using global datasets and validation loss using the Caucasian datasets

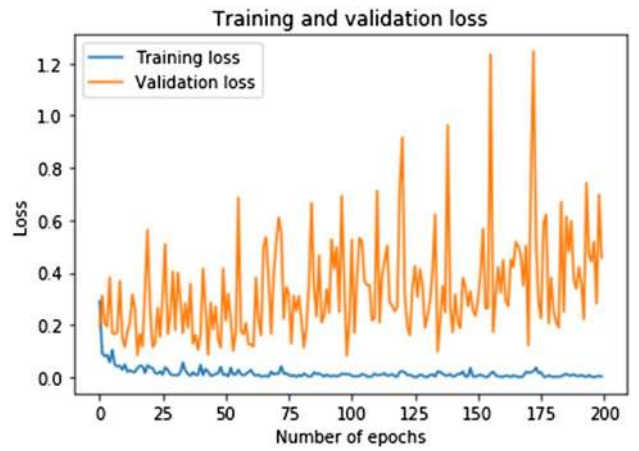

Fig. 29 Showing training loss using global datasets and validation loss using the African datasets

of the racial dataset used at deployment phase. Similarly, Figs. 27, 28 and 29 show similar approach in which training the model was conducted using global dataset while validation using only African dataset.

Table 1 below presents a summary of all the experiments executed in this paper. The results show the machine learning algorithm trained with training dataset containing different racial representations produced the most robust and effective diagnostic system. The terminologies used in the table are Caucasian dataset (Cauc), African dataset (Afri), global dataset (Glo), Reshuffling (Re) and accuracy (Acc).

Table 1 Classification accuracy

Training				Validation			Acc
Cauc	Afri	Glo	Re	Cauc	Afri	Glo	
Y				Y			99.6%
	Y				Y		96.4%
Y					Y		87.5%
	Y			Y			83.4%
		Y				Y	96.4%
		Y	Y			Y	99.5%
		Y		Y			99.3%
		Y			Y		97.1%

4 Conclusion

The study in this paper provide in-depth experiments and analysis of discriminating burns in different ethnic or racial identities achieving a recognition accuracy of up to 99%. The study further provides a baseline for future investigation, specifically in healthcare, on how embedded racial feature representations in the training data provides a robust and flexible diagnostic tool that can be deployable anywhere.

The experimental results reveal that embedded burn features in the training datasets are prone to making a deep learning algorithm more robust and less biased when deployed to a new environment. A model that is well trained on Caucasian datasets performs poorly when testing using African datasets and vice-versa. We show poor the classification accuracy was when a model was trained and test using Caucasian and African datasets, respectively, achieving 87.5%. Similarly, 83.4% of classification accuracy was achieved when the model was trained and testes on African and Caucasian datasets, respectively. But when the model was trained using the global dataset, the classification accuracy of 99.3% and 97.1% was achieved using Caucasian and African datasets respectively during the validation process.

Effective identification accuracy was tested using three databases of burn images (Caucasian, African and global). The prediction accuracy on Caucasian database yielded the best result achieving up to 99.5% accuracy. Conversely, a decrease in performance was observed when the algorithm was trained on African database but outperforming experienced dermatologists' evaluation which indicates an impressive identification performance. Combining the two databases forming a new database (global) ensures future prediction bias is avoided. As such, both image representations from the two ethnic groups are represented. The result achieved state-of-the-art recognition accuracy.

However, the classification results in this paper recorded some misclassification, as shown in the samples depicted in Fig. 29. These misclassifications were attributed to a number of factors ranging from feature similarity between full-thickness burns and some normal (healthy) skin in Caucasian datasets and poor image resolutions, bad illumination during image acquisition in African datasets. Full-thickness burns feature such as whitish and waxy appearances are the major challenge observed. Hence, they were misclassified as healthy skin. Additionally, the presence of unpeeled burn skin in the African datasets was misclassified as healthy skin, as shown in Fig. 30.

One obvious limitation of our proposed pipeline is the lack of comparative analysis of our result with the prior works, and this is due to our inability to get access to the datasets used in the literature, all effort made to get access to the data in the existing works has been futile. Therefore, our work in this paper can serve as a baseline for future investigation as we intend to make our data publicly available in due course. Moreover, extensive data processing is needed to further diminishes the potential misclassification of burns, more specifically the deep burns (such as full-thickness), as healthy to further address challenges of underestimation.

Fig. 30 Sample of misclassified images

Another obvious limitation of our study is lack of experiments to discriminate burn depth which is a critical aspect to make the decision whether the injury is severe enough for surgical intervention, this will be treated in future work.

Acknowledgements Our sincere thanks and appreciation go to the Petroleum Technology Development Fund (PTDF) for the grant to carry out this study (Grant No.: PTDF/ED/PHD/AA/1104/17)

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Grosu-Bularda, A., Andrei, M.-C., Mladin, A. D., Sanda, M. I., Dringa, M.-M., Lunca, D. C., et al. (2019). Periorbital lesions in severely burned patients. *Romanian Journal of Ophthalmology, 63*(1), 38.
2. Rowan, M. P., Cancio, L. C., Elster, E. A., Burmeister, D. M., Rose, L. F., Natesan, S., et al. (2015). Burn wound healing and treatment: Review and advancements. *Critical Care, 19*(1), 243.
3. Charuvila, S., Singh, M., Collins, D., & Jones, I. (2018). A comparative evaluation of spectrophotometric intracutaneous analysis and laser doppler imaging in the assessment of adult and paediatric burn injuries. *Journal of Plastic, Reconstructive & Aesthetic Surgery, 71*, 1015.
4. Shin, J. Y., & Yi, H. S. (2016). Diagnostic accuracy of laser Doppler imaging in burn depth assessment: Systematic review and meta-analysis. *Burns, 42*(7), 1369–1376.
5. Jaspers, M. E., van Haasterecht, L., van Zuijlen, P. P., & Mokkink, L. B. (2019). A systematic review on the quality of measurement techniques for the assessment of burn wound depth or healing potential. *Burns, 45*(2), 261–281.
6. Shah, S. R. M., Velander, J., Perez, M. D., Joseph, L., Mattsson, V., Asan, N. B., Huss, F., & Augustine, R. (2019). Improved sensor for non-invasive assessment of burn injury depth using microwave reflectometry. In *2019 13th European Conference on Antennas and Propagation (EuCAP)*, (pp. 1–5)
7. Abubakar, A., & Ugail, H. (2019) Discrimination of human skin burns using machine learning. In *Intelligent Computing-Proceedings of the Computing Conference*, (pp. 641–647)
8. Gladence, L. M., Karthi, M., & Anu, V. M. (2015). A statistical comparison of logistic regression and different Bayes classification methods for machine learning. *ARPN Journal of Engineering and Applied Sciences, 10*(14), 5947–5953.
9. Tang, T. A., Mhamdi, L., McLernon, D., Zaidi, S. A. R., & Ghogho, M. (2016) Deep learning approach for network intrusion detection in software defined networking. In *2016 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, (pp. 258–263)
10. Zhao, Z., Chen, W., Wu, X., Chen, P. C., & Liu, J. (2017). LSTM network: A deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems, 11*(2), 68–75.
11. Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture, 161*, 272–279.
12. Dong, Y., Liu, Y., & Lian, S. (2016). Automatic age estimation based on deep learning algorithm. *Neurocomputing, 187*, 4–10.
13. Bukar, A. M., & Ugail, H. (2017). Automatic age estimation from facial profile view. *IET Computer Vision*, *11*(8), 650–655.
14. Bukar, A. M., & Ugail, H. (2017). Convnet features for age estimation. In *11th international conference on computer graphics, visualization, computer vision and image processing*.
15. Jilani, S. K., & Driver, S. (2017). Forensic facial recognition. In A. Barbaro (Ed.), *Manual of forensic science* (pp. 111–134). Boca Raton: CRC Press.
16. Lopes, U., & Valiati, J. F. (2017). Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. *Computers in Biology and Medicine, 89*, 135–143.
17. Dong, Y., Jiang, Z., Shen, H., Pan, W. D., Williams, L. A., Reddy, V. V., Benjamin, W. H., & Bryan, A. W. (2017) Evaluations of deep convolutional neural networks for automatic identification of malaria infected cells. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, (pp. 101–104)
18. Sarraf, S., & Tofighi, G. (2016) Classification of Alzheimer's disease using Fmri data and deep learning convolutional neural networks. *arXiv preprint arXiv:1603.08631*
19. Gladence, L. M., Ravi, T., & Dhas, Y. M. (2015). An enhanced method for disease prediction using ordinal classification-APUOC. *Journal of Pure and Applied Microbiology, 9*, 345–349.
20. Chauhan, J., Goswami, R., & Goyal, P. (2018). Using deep learning to classify burnt body parts images for better burns diagnosis. *Sipaim-Miccai Biomedical Workshop* (pp. 25–32). Cham: Springer.
21. Suvarna, M., Toney, G., & Swastik, G. (2017) Classification of scalding burn using image processing methods. In *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, (pp. 1312–1315)
22. Yadav, D., Sharma, A., Singh, M., & Goyal, A. (2019). Feature Extraction Based Machine Learning for Human Burn Diagnosis From Burn Images. *IEEE Journal of Translational Engineering in Health and Medicine, 7*, 1–7.
23. Abubakar, A., Ugail, H., & Bukar, A. M. (2019) Can machine learning be used to discriminate between burns and pressure ulcer? In *Proceedings of SAI Intelligent Systems Conference*, (pp. 870–880)
24. Cirillo, M. D., Mirdell, R., Sjöberg, F., & Pham, T. D. (2019). Time-independent prediction of burn depth using deep convolutional neural networks. *Journal of Burn Care & Research., 40*, 857.
25. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.
26. Ferreira, A., & Giraldi, G. (2017). Convolutional neural network approaches to granite tiles classification. *Expert Systems with Applications, 84*, 1–11.
27. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012) Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, (pp. 1097–1105)
28. Simonyan, K., & Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*

29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 1–9)
30. He, K., Zhang, X., Ren, S., & Sun, J. (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 770–778)
31. Abubakar, A., Ugail, H., Bukar, A. M., & Smith, K. M. (2019). Discrimination of healthy skin, superficial epidermal burns, and full-thickness burns from 2D-colored images using machine learning. *Data Science* (pp. 201–223). Boca Raton: CRC Press.
32. Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2019) Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *arXiv preprint arXiv:1912.07398*.