Journal of Medical and Biological Engineering (2020) 40:923–933
https://doi.org/10.1007/s40846-020-00574-z

ORIGINAL ARTICLE

Burns Depth Assessment Using Deep Learning Features

Aliyu Abubakar[1,3] · Hassan Ugail[1] · Kirsty M. Smith[2,4] · Ali Maina Bukar[1] · Ali Elmahmudi[1]

Received: 5 June 2020 / Accepted: 8 October 2020 / Published online: 16 October 2020

Abstract

Purpose Burns depth evaluation is a lifesaving task and very challenging that requires objective techniques to accomplish. While the visual assessment is the most commonly used by surgeons, its accuracy reliability ranges between 60 and 80% and subjective that lacks any standard guideline. Currently, the only standard adjunct to clinical evaluation of burn depth is Laser Doppler Imaging (LDI) which measures microcirculation within the dermal tissue, providing the burns potential healing time which correspond to the depth of the injury achieving up to 100% accuracy. However, the use of LDI is limited due to many factors including high affordability and diagnostic costs, its accuracy is affected by movement which makes it difficult to assess paediatric patients, high level of human expertise is required to operate the device, and 100% accuracy possible after 72 h. These shortfalls necessitate the need for objective and affordable technique.

Method In this study, we leverage the use of deep transfer learning technique using two pretrained models ResNet50 and VGG16 for the extraction of image patterns (ResFeat50 and VggFeat16) from a a burn dataset of 2080 RGB images which composed of healthy skin, first degree, second degree and third-degree burns evenly distributed. We then use One-versus-One Support Vector Machines (SVM) for multi-class prediction and was trained using 10-folds cross validation to achieve optimum trade-off between bias and variance.

Results The proposed approach yields maximum prediction accuracy of 95.43% using *ResFeat50* and 85.67% using *VggFeat16*. The average recall, precision and F1-score are 95.50%, 95.50%, 95.50% and 85.75%, 86.25%, 85.75% for both *ResFeat50* and *VggFeat16* respectively.

Conclusion The proposed pipeline achieved a state-of-the-art prediction accuracy and interestingly indicates that decision can be made in less than a minute whether the injury requires surgical intervention such as skin grafting or not.

Keywords Skin burns · Burn depths · Deep learning · Features · SVM · Classification

1 Introduction

Skin is the largest body organ constituting ~ 1.5–2.0 m^2 for an average adult [1]. It serves as a defensive shield against foreign intruders, helps in thermoregulation, prevents loss of body fluid via evaporative, and helps significantly in the production of vitamin D. Skin is composed of three layers: epidermis, dermis and hypodermis. The epidermis is the outermost layers that interface the external environment while dermis sits between epidermis and hypodermis. These skin layers, combined together, provide the aforementioned functionalities. However, skin injuries such burns disrupt such barrier thereby subjecting individuals to high risk of infections and in extreme cases loss of live. Burns injuries are caused by several mechanisms such as thermal, electrical, radiation and chemical [2]. Burns that affect epidermal layer are referred to as superficial or first-degree burns, and the common example is sunburn which can heal with no medical intervention within seven days due to proliferation and differentiation of keratinocytes from basal epithelial cells [1]. Deep burns such as second-degree and third-degree burns are distinguishable from epidermal burns by their characteristics (pain, capillary refill and colour-red/pink, white).

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

1 Centre for Visual Computing, Faculty of Engineering and Informatics, University of Bradford, Bradford, UK

2 Plastic Surgery and Burns Research Unit, Centre for Skin Sciences, University of Bradford, Bradford BD7 1DP, UK

3 Department of Computer Science, Faculty of Science, Gombe State University, Gombe 760214, Nigeria

4 Bradford Teaching Hospitals NHS Foundation Trust, Bradford, West Yorkshire, UK

Second-degree burn includes superficial partial-thickness (SPT) burns and deep partial-thickness (DPT) burns. SPT are characterized by pain and capillary refill and involves both epidermis and papillary dermis while DPT burns extend to reticular dermis and adnexal structures. Third-degree burns (also referred to as full-thickness burns) affect all the epidermal and dermal layers and extend to subcutaneous adipose tissue, muscles and bones.

Patients with a second degree burn, specifically DPT, and third-degree burn require immediate and effective assessment for early recovery. These injuries take substantial lengthy hospitalization and have the high risk of subjecting patients to hypertrophic scars (HTS). However, burn depth assessment has been challenging task for clinicians. Assessment by experienced clinicians is highly subjective due to lack of standard guideline with accuracy ranging between 60 and 80% [1, 3], which prompted the need for a better alternative modality.

Other objective techniques have been proposed for burns depth assessment. Knowing the depth of burn injury gives crucial information regarding the expected recovery time (healing time). These proposed objective methods include the use of Laser speckle imaging, spatial frequency domain imaging and laser doppler imaging [4]. Laser speckle imaging (LSI) has the capability to assess the perfusion rate over a wide area, easy to interpret the produced perfusion map by clinicians. A map area that shows high perfusion rate simply means vasculature is undamaged and the burn area is likely to heal without any medical intervention while low perfusion rate basically means damaged tissue and may require quick surgical intervention [4]. However, the accuracy is time-dependent with optimum performance at 48–72 h after burn occurrence while inaccurate and inconclusive before 48 h [5]. LDI is a non-invasive tool first used for burns examinations by Niazi et al. [6] that scans tissue surface using monochromatic laser beam and gets reflected by moving blood cells. The extent of reflection correlates with the severity of the damaged tissue, where high reflection corresponding to the high perfusion rate and indicates very shallow burns while deeper burn wounds are determined by low perfusion rate (with low reflection) because there is lesser blood circulation as a result of blood vessels been damaged by the burn injury. LDI remains the prominent tool and widely accepted for burns examination today with additional advantage of scanning wide area of up to 50 cm by 50 cm, unfortunately the cost of the equipment is high with an estimated cost of £50,000 [7–9], it is cumbersome, it requires high expertise to operate, it is slow where a scan takes up to 1 or 2 min. Laser speckle contrast imaging (LSCI) is a recent objective technique for measuring microcirculation non-invasively that shorten the scanning time to about 200–1000 ms compared to LDI [8, 10], and less sensitive to patient movement artifacts. Despite its performance and advantages over LDI, its usage has limited application on burn evaluation and achieve accuracy of approximately 95% from day 3 after injury [10–12].

Towards the end, this proposed research provides alternative burn depth evaluation using deep learning features to objectively predict those burns that that require surgical intervention and those that do not. In summary, the contributions of this research are outlined below:

- We introduce **ResFeat50** and **VggFeat16**, image features extracted from ImageNet pretrained models, ResNet50 and VGG16, respectively, to predict human skin burns healing times and Support Vector Machines as a predictor.
- We provide an in-depth analysis regarding features with strong discriminatory patterns and made based on their robustness and computational time comparison between ***ResFeat50*** and ***VggFeat16.***
- We provide performance comparison of our proposed study with the existing published works. Our approach and results achieved a significant performance improvement

The rest of the paper is organized as follows: in the next section, we briefly discuss related works and Sect. 3 presents methodology. In Sect. 4, we present experimental results, Sect. 5 presents discussion of the results and Sect. 6 concludes the paper.

2 Literature

The use of Convolutional Neural Networks (CNN) for classification tasks has widely been adopted in different application domains such as face recognition [13, 14] and disease detection [15]. Their adoption was due to their capability to capture rich generic discriminatory features at different levels. It was proposed in a study by authors in [16] to discriminate whether a given human skin image is burnt or healthy. This was facilitated using pretrained CNN features, specifically ResNet101 model was used, due to deficient datasets. The datasets are all RGB images and pre-processed by resizing them to a standard input size of ResNet101 model. Thereafter, the extracted features were fed into support vector machines and trained using tenfold cross validation. This approach recorded a near perfect classification accuracy of 99.5%.

Another study referenced [17] lamented a challenge if the deep learning model is to be trained from scratch using limited dataset. Alternatively, the study opted for transfer learning for deep feature extraction, an approach known as off-the-shelf feature extraction. Two pretrained residual network models, ResNet101 and ResNet152, were used for

features extraction to discriminate between burn wounds and pressure ulcer injuries and support vector machines was trained for the classification via the use of tenfold cross validation. ResNet152 features proved to have more strong discriminatory patterns from the images in which support vector machines recorded 99.9% accuracy.

Study referenced [18] proposed another binary classification of burns using deep neural network features and support vector machinesthe. In this study, three deep CNN models were used; two of the models (VGG16 and VGG19) were training on ImageNet database to categorize 1000 different objects and the other model was trained to recognize human faces (VGGFace). In nutshell, all the three pretrained CNN model were used for feature extraction and then support vector machines as classifier,a. Results show that 98.75% and 97.56% using VGG16 and VGG19 features respectively, while achieving 95.20% on VGGFace. Finally, the authors lamented that high accuracy recorded by ImageNet models can be attributed to the fact that the weights of those models were able to learn from a diverse representation of features.

Similarly, the study referenced in [19] proposed a binary classification of burns and healthy skin using fine-tuning approach. A pre-trained ImageNet deep learning model (i.e. ResNet50) was adopted and modified the top layers. Two different datasets from two ethnicities were used; Africans and Caucasians. The dense layers of the ResNet50 model were removed and replaced with new layers, these layers were then trained using features from the base layers of the ResNet50. Recognition accuracy of 97.1% on African images and recorded classification accuracy of 99.3% on Caucasian images. The authors attributed lack of good recognition accuracy for the African subjects due to poor quality of the images.

A study by [20] proposed an automated diagnostic process to classify burn wounds. In this study, discrimination of burn images and injured skin (pressure ulcer and skin bruises) was conducted. The study invoked two transfer learning approaches due to insufficient datasets; fine-tuning which involves modifying top layers of deep learning model, and on the other hand training support vector machines using features extracted by the pre-trained deep learning model. Three pre-trained models, including ResNet50 with 50 stacked convolution layers, ResNet101 contained 101 stacked convolution layers and ResNet152 containing 152 stacked convolution layers were employed and compared. In the end, Training support vector machines with ResNet152 features recorded the best classification accuracy of 99.96% with area under the curve (AUC) of 99.99%. Fine-tuning requires considerable database size.

For burn depth recognition, few numbers of studies used machine learning techniques. For example, SPT, DPT and full-thickness burns were classified using machine learning in a study referenced [21]. Total of 164 images acquired, and all were converted into L*a*b* colour space. Prior to feature extraction, relevant regions of interest were segmented, discrete wavelet transforms (DWT) was used to specifically extract texture features and principal component analysis (PCA) was additionally used to reduce the dimensionality of the features. The best classification accuracy achieved was via the use of simple logistic regression which recorded 73.2%.

Discriminating burns depth using machine learning was also reported in a study by [22]. The aim is to provide a reliable diagnostic technique to deduce whether a sustained burn injury requires surgical intervention or not because early determination of right treatment choice can shorten the healing time. 74 RGB images were transformed into L*a*b* colour space and extract certain features: hue, hog, chroma, kurtosis and skewness. Thereafter, support vector machine was trained and achieved a classification accuracy of 82.43%, with precision, recall and F1-score of 82%, 88% and 85% respectively.

In another study [23] using 450 burn images, the study was proposed to discriminate different categories of burn depths. These images were transformed into YCbCr colour space and resized into 120×120 pixels. In each category of burn wounds (first, second and third degrees) there are 150 images representing each category. Thereafter, the authors segmented regions of interest, and used deep CNN architecture to classify the images based on their specific feature of interest (colour and texture) with classification accuracy of 79.4%.

3 Materials and Methods

In this section, data acquisition and preparation are presented. Proposed system architecture for the discrimination of the classes of burn injuries is quantitatively explained and presented.

3.1 Data Acquisition

In this study, we gathered the datasets used for the experiment from both internet search and hospital. Those obtained from the internet are mainly first degree (1DB) burns (mostly sunburn images), and these are injuries that can heal in less than 7 days on their own without any complicated assessment. While the deeper wounds which include Second degree burns and third-degree burns were acquired ethically from Bradford Teaching Hospitals United Kingdom. These images are from different parts of the body, some from upper limbs, lower limbs, back, face and neck.

3.2 Ground Truth Definition

In order to train a machine learning algorithm, specifically in supervised approach, there is an absolute need to annotate the available data effectively by specialists. This annotation process was facilitated using LDI device to label the burn depth regions effectively. LDI measures disruption of the blood flow in the blood vessels and the speed of the blood flow indicates how deep the burn wound is. High reduction of dermal blood flow is observed if the wound is deeper due to the damaged blood vessels [24]. LDI produces a colour map of the wound; red/yellow areas indicating high perfusion rate particularly for superficial epidermal and superficial dermal burns, green indicating low(moderate) perfusion rate for deep dermal burn, and blue colour indicating very low perfusion rate for full thickness burn. The different burn depth were labelled by experts after patients were assessed using the LDI device. Thereafter, regions of interest were extracted out corresponding to the following categories: second degree burns (2DB) that heals between 14 to 21 days and third-degree burns (3DB) that heal after 21 days. The definition of the ground truth made the specialist is displayed in Table 1 and sample of the dataset is depicted in Fig. 1. Note that, 2DB and 3DB images are heterogeneous, which means some pixels in 2DB contain 3DB feature and some in 3DB contain 2DB features and its very fiddficult, if not impossible, to crop out each patch. Doing so will result to a very smaller image with poor resolution. In order to deal with this situation, we established a simple criteria. This criteria states that:

- a given image is 2DB if such depth constitute not less than 80% of the total depth area
- a given image is 3DB if such depth constitute not less than 80% of the total depth area

3.3 Data Augmentation

Deep CNN are data-hungry algorithms that require enormous data to be trained and learn from. Most at times these data are not sufficiently available particularly in medical field due to either privacy concern or lack of experts for the data annotation. One of the available and the most applied method to overcome data deficiency is data augmentation [25]. Data augmentation involves different processes

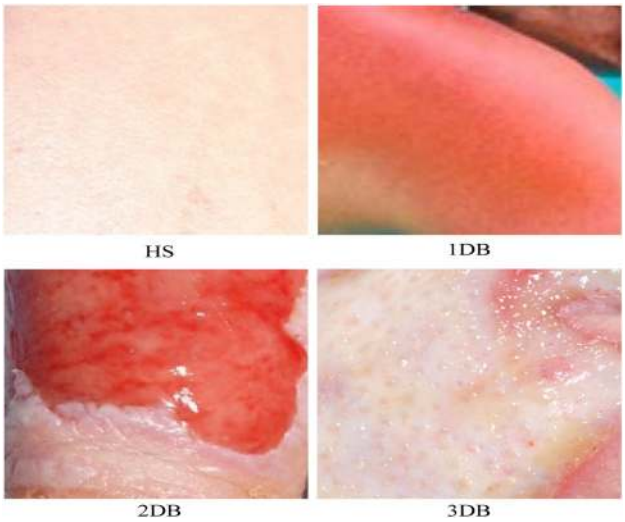

Fig. 1 Showing dataset samples

of transforming original data to produce new instances of same nature with different spatial orientations. These processes include rotation, random cropping, zooming, channel shifting, whitening and flipping. The size of the database was enlarged using two of such transformation processes (rotation and flipping). Rotation involves rotating the images with various degrees such as 45^{o}, -45^{o} and 75^{o}. Flipping mirrored or flipped the given image vertically or horizontally. The information of the enlaged database is presented in Table 2.

3.4 Choice of a Feature Extractor

Deep CNN are fee-forward neural network containing multiple hidden layers interconnected with each other. Training deep CNN requires repetitive adjustment of parameters such as weights, biases and activations in order to produce a satisfying output.

Generally, CNN can be trained in three different ways [26]: training from scratch which requires a lot of hyperparameters tweaking. Hyperparameters tweaking includes adjusting the CNN topology, how the neurons in the network are interconnected, number of network layers, number of neurons in each layer, the activation function to be used and a lot more. It is also important to note that, training a CNN from scratch requires enormous data which is often a

Table 1 Defined ground truth datasets

Depth	< 7 days	14–20 days	> 21 days
1DB	163	0	0
2DB	0	450	0
3DB	0	0	130

Table 2 Augmented datasets

Depth	< 7 days	14–20 days	> 21 days
1DB	520	0	0
2DB	0	520	0
3DB	0	0	520

very challenging task. The second way of deploying CNN is fine-tuning which involves transferring the weights of learned layers from an existing network to a new network. Thirdly, CNN can be used as off-the-shelf feature extractor so that strong discriminatory features can be extracted and subsequently used those features to train a different machine learning classifier. Due to simplicity, lack of enough data and computational resources to train CNN from scratch, we opted to use the latter approach for feature extraction. There are several pre-trained CNN models available for off-the-shelf feature extraction such as AlexNet, GoogleNet, VGG-Net and ResNet.

Therefore, two pre-trained ImageNet CNNs (VGG16 and ResNet50) are used for feature extraction in this study. The choice was inspired by the fact that CNN models trained on multiple data categories have strong generic information that can be used on the fly for image feature representation [18, 20, 27].

3.4.1 Image Pre-processing

Prior to feature extraction, all images must conform to standard input requirement of the feature extractor, as such we made sure they are resized to a standard size corresponding to the input specification of the feature extraction model. Both ResNet50 and VGG16 has same input size configuration of accepting input data of size 224 × 224, and this is the only pre-processing performed before passing the images for pattern extraction.

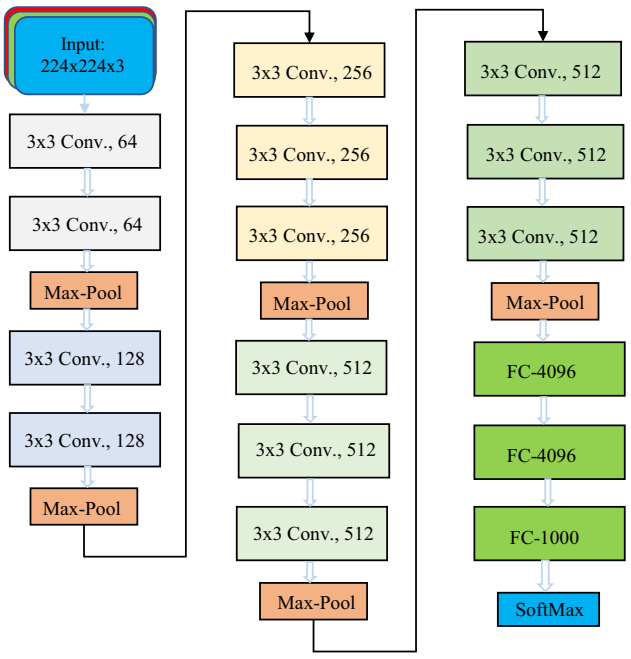

Fig. 2 Illustration of VGG16 model architecture

3.4.2 Feature Extraction Using VGG16

VGG16 was developed by Visual Geometry Group research teams at Oxford University and was trained on ImageNet database in 2014 [28, 29]. VGG16 has a total of 37 layers; 13 of them are convolution layers as illustrated in Fig. 2, and the remaining layers consist of mixed of pooling, activation and fully connected layers. VGG used smaller filter size of 3 × 3 throughout the network and has proved to be computationally efficient compared to large filter size used in AlexNet and is considerably deep to learn more complex patterns. Since CNN layers learn different types of features as the data propagates down through the network; the lower layers learn low-level features while the deeper layers learn high-level or more abstract features, the first fully connected layer was used to collect the generic features denoted as ***VggFeat16***..

3.4.3 Feature Extraction Using ResNet50

ResNet50 is one of the Residual Network (ResNet) models by Microsoft Research Asia, the winner of ImageNet Large Scale Visual Recognition Challenge in 2015 [30]. The model is stacked with 50 convolution layers, including a fully connected layer with 1000 neurons. Though increasing the network depth to a certain limit leads to degradation of accuracy and overfitting problems, ResNet has overcome these problems via the use of identity mapping as illustrated in Fig. 3. Instead of learning direct mapping x*toy*, *y* is reframed into $y = \sigma(f(x) + x)$ where σ is a non-linearity function, this enables it to grow deeper achieved outstanding performance [31]. This impressive breakthrough inspired the idea of using ResNet50 to pull out abstract image patterns denoted as ***ResFeat50 in this paper. ResFeat50*** were collected at the last year before the classification year.

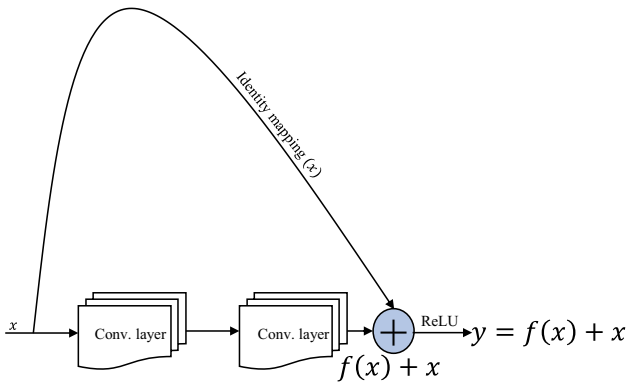

Fig. 3 Illustration of ResNet residual block

3.5 Feature Rescaling

Most at times data are composed of attributes with varying scales, when such data are rescaled, machine learning algorithms benefit greatly and perform remarkably well, and it is also useful for optimization algorithms [32]. As such, after the features extraction we applied rescaling process, often referred as normalization and the features are rescaled into the range of 0 and 1.

3.6 Choice of a Classifier (Predictor)

We used linear support vector machines (SVM), a supervised machine learning algorithm widely used for binary classification [33]. When the number of observations and their corresponding labels are given, SVM works by finding a separating boundary (optimum separating hyperplane) in the given feature space, so that each instance is placed in a different semi-space and trying to maximize the distance separating them thereby minimizing misclassification errors [34, 35].

The SVM can be represented as the linear combination of features designated as x, multiplied by weights ω as presented in Eq. (1):

$$f(x) = \omega^T + b = 0 \tag{1}$$

where ω and x ϵ $\boldsymbol{R}^d$, d is the size of the space, and b is the noise or bias. Depending on which side the samples are located with respect to the hyperplane, Eq. (2) and Eq. (3) define the scenarios for the two classes.

$$\omega^T x_i + b > 0, \; for \; y_i = 1, i = 1, \ldots, n \tag{2}$$

$$\omega^T x_i + b < 0, \; for \; y_i = -1, i = 1, \ldots, n \tag{3}$$

Interestingly, SVM can also be tweaked to solve problem of more than two classes. one of the methods of classifying multiple classes using SVM is One-versus-One (OVO) [36] classification strategy which we adopted in this study. Using OVO, the number of classes (N_c) are broken down into multiple binary classification problem. When dealing with multiple classes, the number of binary classifiers produced using OVO strategy is defined by Eq. (4):

$$Number\ of\ binary\ problems = \frac{N_c\left(N_c - 1\right)}{2} \tag{4}$$

Evaluating SVM performance was carried out using one of the most famous evaluation techniques, a cross-validation (CV) [37]. This technique works by splitting the whole datasets into K equal folds. K-1 folds are then used to train the SVM and the withheld fold used for testing. The process is repeated until each fold out of the K-folds gets chance to be used as testing split. At the end of the runs, the accuracy of the SVM is obtained by averaging the performance measures across all folds. The most commonly used values for K are 3,5,7, and 10, in this study we used K = 10. One of the benefits of training a classifier using CV is to mitigate overfitting problem.

4 Experimental Results

In this section, two experiments conducted to discriminate the four classes of images is presented. SVM is trained using the two deep image patterns (***ResFeat50*** and ***VggFeat16)***. Discriminating the different degree of burns here will render vital information to burn surgeons and other health practitioners whether there shall be a need for a patient to undergo surgical intervention or just wound dressing. Predicting burn image as 1DB indicates a burn that can heals in first seven days after injury and does not require surgery. Predicting burn as 2DB indicates an burn that may require

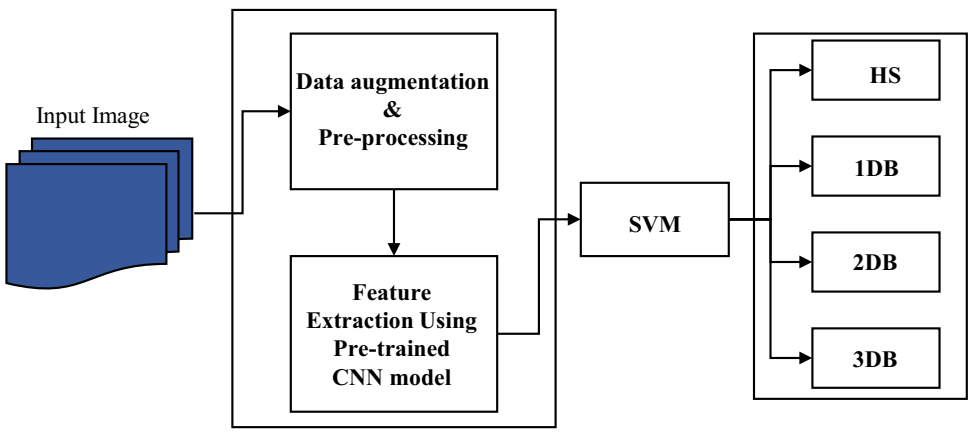

Fig. 4 Experimental set-up

Table 3 Predicted output using **ResFeat50**

Target				
Predicted	HS	1DB	2DB	3DB
HS	499	0	6	15
1DB	0	503	11	6
2DB	5	27	480	8
3DB	4	2	11	503

Table 4 Prediction output using **VggFeat16**

Target				
Predicted	HS	1DB	2DB	3DB
HS	482	18	15	5
1DB	15	463	13	29
2DB	11	32	379	98
3DB	7	9	46	458

surgery and takes a bit long time to heal, normally can take up to 2–3 weeks, while a burn predicted to be a 3DB can take more than three weeks to heal and requires surgery the most. Note that, early assessment can help in shortening the healing time if the necessary intervention is provided Fig. 4 depicts the experimental set-up.

We then used an error matrix [35], which is a multi-dimensional table use for visualizing classifier performance. It shows a combination of actual values and predicted values, this enables to determine whether the classifier has predicted individual instances belonging to each class accurately or it has performed erroneously. Confusion matrix displayed in Tables 3 and 4 show the predicted outputs using ***ResFeat50*** and ***VggFeat16*** respectively.

Target (actual) classes are represented as columns as shown in Table 3 which presents SVM's performance using ***ResFeat50***, while the rows values represent predicted classes by the classifer. Out of the 520 healthy skin (HS) images, 499 were classified accurately, 6 were misclassified as 2DB, 15 were misclassified as 3DB while none was misclassified as 1DB. Out of the 520 1DB images, none was misclassified as HS, 11 were misclassified as 2DB images, 6 were misclassified as 3DB and 503 were accurately classified as 2DB. For the 2DB images, 5 were misclassified as HS, 27 misclassified as 1DB, 8 misclassified as 3DB while 480 were accurately classified. Lastly, out of the 520 3DB images, 503 were accurately classified, 11 misclassified as 2DB, 2 misclassified as 1DB and 4 misclassified as HS.

Similarly, Table 4 presents the classification output of SVM using ***VggFeat16.*** Out of the 520 HS images, 482 were classified accurately, 15 were misclassified as 2DB, 5 were misclassified as 3DB while 18 were misclassified as 1DB. Out of the 520 1DB images, 15 were misclassified as HS, 13 were misclassified as 2DB images, 29 were misclassified as 3DB and 468 were accurately classified. For the 2DB images, 11 were misclassified as HS, 32 misclassified as 1DB, 98 were misclassified as 3DB while 379 were accurately classified. Lastly, out of the 520 3DB images, 458 were accurately classified, 46 misclassified as 2DB, 9 misclassified as 1DB and 7 misclassified as HS.

Comparateively, ***ResFeat50*** contains more discriminatory features which led to the SVM performance more effective on those features. In general, 95 misclassifications by the classifier on ***ResFeat50*** while on VggFeat15 there are 298 misclassifications. About 62 out of 520 patients with 3DB may be subjected to unnecessary delay if ***VggFeat16*** were used to assessment, and 17 out of 520 with 3DB could perhaps be subjected to unnecessary delay if ***ResFeat50*** was used for the assessment.

4.1 Performance Evaluation Metrics

In order to evaluate the performance of the prediction, there is need to interpret the values obtianed in the Tables 3 and 4. These evaluation measuresare based on the following parameters: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) [38, 39]. Accuracy is on of those metrices that gives the general performance of the classification.

4.1.1 Accuracy

This determines the classifier's correctness in predicting actual classes correctly as the Eq. 5 provided. This gives the accurate prediction of the whole classifier.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

4.1.2 Recall

This measure is the ability of the classifier to predict each individal class correctly. Recall (or sensitivity) gives the accurate prediction of individual class by the classifier. Equation (6) provdes mathematical formula for determining recall:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

In this scenario, FN can include all predictions made by classifier to other classes belongng to a particular class in question and TP stands for the correct prediction of that class. This can also be interpreted as

Table 5 Classification accuracy

Features	Accuracy (%)	Time(sec)
VggFeat16	85.67	147.9
ResFeat50	95.43	39.1

Table 6 Performance metrics using VggFeat16

	Precision	Recall	F1-score
HS	0.94	0.93	0.93
1DB	0.89	0.89	0.89
2DB	0.84	0.73	0.78
3DB	0.78	0.88	0.83

Table 7 Performance metrics using ResFeat50

	Precision	Recall	F1-score
HS	0.98	0.96	0.97
1DB	0.95	0.97	0.96
2DB	0.94	0.92	0.93
3DB	0.95	0.97	0.96

$$Recall = \frac{True\ positives}{Total\ actual\ positives}$$

4.1.3 Precision

this measure determines the fractions of relevant or true instances predicted by the classifier, and its mathematically expressed in Eq. (7):

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

So, the above equation can simply be interpreted as

$$Precision = \frac{True\ positives}{Total\ predicted\ positives}$$

4.1.4 F1-score

This metric combines both recall and precision and presents the two as a single measure. It is simply a harmonic mean of the two metric measures (recall and precision) as provided in Eq. (8).

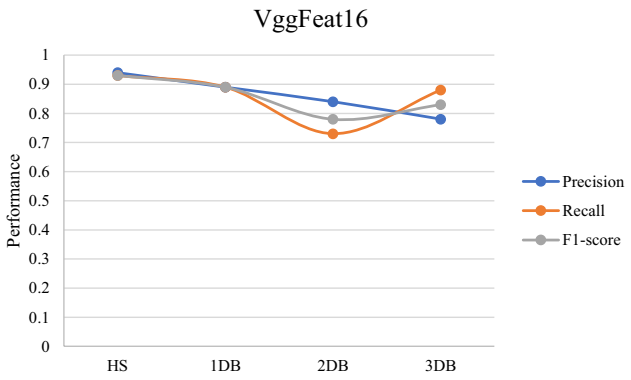

Fig. 5 Showing comparison of performance evaluation measures using VggFeat16

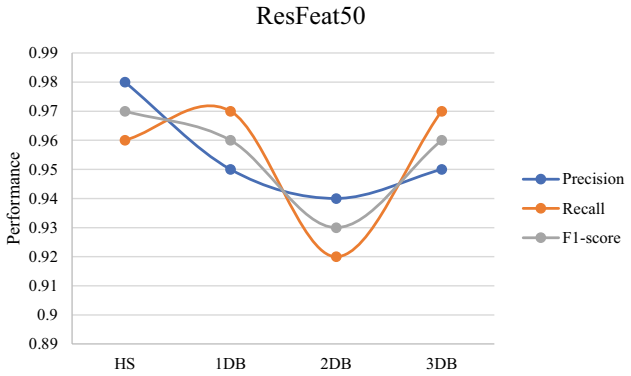

Fig. 6 Showing comparison of performance evaluation measures using ResFeat50

$$F1 - score = \left(\frac{2}{Recall^{-1} + Precision^{-1}}\right) \tag{8}$$

Table 5 ptovides the overall accuracy of the two different experiments, SVM predicted more accuractely using *ResFeat50* achieving 95.43% than *VggFeat16* with 85.67%. Tables 6 and 7 provide performance evaluation values using both *VggFeat16* and *ResFeat50* respectively. SVM performed better with *ResFeat50* than *VggFeat16*.

The results in Tables 6 and 7 are depicted in Figs. 5 and 6 respectively for good visualization. Both Fig. 5 and Fig. 6 show that 2DB injuries are also difficult to assess using machine learning techniques, but the perormance is impreesive and better than experienced health specialist.

5 Discussion of Results

Long hospitalization (LH) is an unpleasnt experience that subjects both patients and their families which can leads to further burn management complications such as increase in

hospital cost. LH can also be attributed to treatment delay due to lack of adequate objective assessment techniques and lack of access to proximity burn centres. Our proposed pipeline has successfully achieved an impressive predcition accuracy of those burn wounds that can heal within a week with no medical intervention required, within two to three weeks and those that can heal in more than three weeks.

We obtained impressive results using *VggFeat16*, the classifier recorded recall of 93%, 89%, 73% and 88% for HS, burns healing with no required hospital management (1DB), burns healing withi two to three weeks (2DB) and burns taking longer time to heal (3DB) which in extreme cases will require skin grafting respectively. Similarly, the precision achieved by the classifier on HS is 94%, 89% for 1DB, 84% for 2DB and 78% for 3DB while f1-score on each of the class precited are 93% for HS, 89% for 1DB, 78% for 2DB and 83% for 3DB. The prediction was succesfully carried out in approximately 148 s (less than 3 min) which ultimately suggests that using machine learning techniques as aiding tools to evaluate burn wounds can facilitate decision-making as early as possible thereby minimising chances of subjecting patients to long hospital delay.

Similarly, using *ResFeat50*, the classifier's prediction is more precise achieving recall of 96% for HS compared to 93% using *VggFeat16*, 97% for 1DB compared to 89% using *VggFeat16*, 92% for 2DB compared to 73% using *VggFeat16* and 97% for 3DB compared to 88% using *VggFeat16*. Similarly, precision recorded by the classifier using *ResFeat50* has surpassed the precision recprded by the classifier using VggFeat16 as presented in Table 7. Moreover, F1-score using *ResFeat50* are 97%, 96%, 93% and 96% for HS, 1DB, 2DB and 3DB respectively. In orderf to find out the trade-off between accuracy and computational time, the classifier is more accurate and efficient using *ResFeat50* with computational time of approximately 39 s (less than a minute).

The *VggFeat16* has 4096 feature vectors while *ResFeat50* has 2048 feature vectors which were used in training the classifier bu the robustness of the classifier is more efficient using *ResFeat50* than *VggFeat16* despite the later having more feature vectors. This simply indicates that *ResFeat50* carries strong dicriminatory features than *VggFeat16*. It is also worth noting that *ResNet50* has perhaps contains more discriminating attributes than *VggFeat16* due to number layers for the feature extraction. Figure 5 provides the F1-score performance comparison of the two feature set predicted by the classifier.

Studies in the literature used very deficient databases, the authors in [21] reported 73.2% accuracy on datasets of 164 images and all images were in L*a*b* colour space with the application of PCA for dimensionality reduction on texture features. Study in [22] reported overall accuracy of 82.43% using 74 images in L*a*b* colour space, another study in [23] reported a discriminatory accuracy of 79.4% using colour and texture features and DCNN as a classifier on a database of 450 images. In this proposed study, we used 1560 RGB burn images along with 520 healthy skin images thereby achieving state-of-the-art discriminatory accuracy of 95.43%. Though the comparison might not be realistic since in this study, a completely different database was used in this study because access to the databases used by studies in the literature was unsuccessful (Fig. 7).

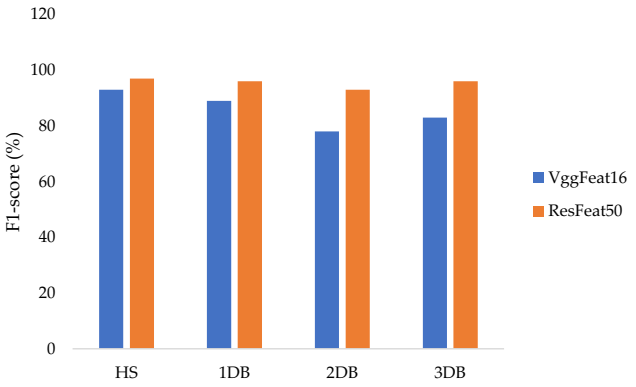

Fig. 7 F1-score comparison of the SVM prediction using *VggFeat16* and *ResFeat50*

6 Conclusion

This study provides an automated process for predicting burns healing times which by similitude refers to burn depths prediction using machine learning. We evaluated the performance of using deep off-the-shelf features via the use of One-versus-One SVM so solve a multi-class problem, specifically predicting burn depths. The useful discriminatory features were extracted from the images using pre-trained ResNet50 and VGG16 models.

Proposed approach achieved 85.67% prediction accuracy on VGG16 features (*VggFeat16*) while ResNet50 features (*ResFeat50*) recorded a maximum and state-of-the-art prediction accuracy of 95.43%. The result indicates aiding burn management in hospitals using the proposed method has the potential of minimizing both under-estimation and over-estimation which is heavily associated with traditional approach (clinical evaluation). Our result obviously shows that chances of under-estimating those burn wounds that may require surgery or skin grafting can be minimized significantly, while those that do not require surgery may not be subjected to unnecessary management thereby incurring additional complications and cost.

This study has recorded some misclassifications that mostly ocuured between 2 and 3DB. This is attributed to the heterogeneity of the datasets. Most of the images contain mix of superficial dermal and full-thickness wound. Similarly, misclassification between HS and 3DB is due to similarity of some instances, some 3DB images look white and leathery. Moreover, poor illumaination of some images contributed to the classification error involving 1DB.

Our result is not without limitation, it's obvious that there is still room for improving the efficacy to minimise the classification errors further using a larger sample size and to specifically discriminate between the two categories of burns that made up of 2DB. 2DB is composed of SPT and DPT burns. In most cases, deep partial thickness burns are the actual category of dermal burn wounds that require surgery. Furthermore, estimating the affected body surface area is clinically important, such will provide a useful hint to determine the size of the skin needed perhaps if skin grafting is inevitable.

Acknowledgement We thank the Petroleum Technology Development Fund (PTDF) Nigeria for funding this Research (Grand Number: PTDF/ED/PHD/AA/1104/17

References

1. Stone, I., et al. (2018). Advancements in regenerative strategies through the continuum of burn care. *Frontiers in Pharmacology, 9,* 672.
2. Pencle, F. J., Zulfiqar, H., & Waseem, M. (2019). *First degree burn*. Treasure Island: StatPearls.
3. Mirdell, R. (2019). *Blood flow dynamics in burns*. Linköping: Linköping University Electronic Press.
4. Ponticorvo, A., et al. (2020). *Spatial Frequency Domain Imaging (SFDI) of clinical burns: A case report.* Burns Open.
5. Ponticorvo, A., et al. (2019). Evaluating clinical observation versus spatial frequency domain imaging (SFDI), laser speckle imaging (LSI) and thermal imaging for the assessment of burn depth. *Burns, 45*(2), 450–460.
6. Niazi, Z., et al. (1993). New laser Doppler scanner, a valuable adjunct in burn depth assessment. *Burns, 19*(6), 485–489.
7. Burke-Smith, A., Collier, J., & Jones, I. (2015). A comparison of non-invasive imaging modalities: Infrared thermography, spectrophotometric intracutaneous analysis and laser Doppler imaging for the assessment of adult burns. *Burns, 41*(8), 1695–1707.
8. Mirdell, R., et al. (2020). Using blood flow pulsatility to improve the accuracy of laser speckle contrast imaging in the assessment of burns. *Burns*. https://doi.org/10.1016/j.burns.2020.03.008
9. Hoeksema, H., et al. (2014). A new, fast LDI for assessment of burns: A multi-centre clinical evaluation. *Burns, 40*(7), 1274–1282.
10. Mirdell, R., et al. (2018). Accuracy of laser speckle contrast imaging in the assessment of pediatric scald wounds. *Burns, 44*(1), 90–98.
11. Heeman, W., et al. (2019). Clinical applications of laser speckle contrast imaging: A review. *Journal of Biomedical Optics, 24*(8), 080901.
12. Mirdell, R., et al. (2016). Microvascular blood flow in scalds in children and its relation to duration of wound healing: A study using laser speckle contrast imaging. *Burns, 42*(3), 648–654.
13. Jilani, S. K., et al. (2017). *A machine learning approach for ethnic classification: The British Pakistani face*. In: *2017 international conference on cyberworlds (CW)*. 2017. IEEE.
14. Elmahmudi, A., & Ugail, H. (2018). *Experiments on deep face recognition using partial faces*. In: *2018 international conference on cyberworlds (CW)*. 2018. IEEE.
15. Polat, K., & Koc, K. O. (2020). Detection of skin diseases from dermoscopy image using the combination of convolutional neural network and one-versus-all. *Journal of Artificial Intelligence and Systems, 2*(1), 80–97.
16. Abubakar, A., & Ugail, H. (2019). *Discrimination of human skin burns using machine learning*. Cham: Springer.
17. Abubakar, A., Ugail, H., & Bukar, A. M. (2019a). Can machine learning be used to discriminate between burns and pressure ulcer? *Proceedings of SAI intelligent systems conference*. Berlin: Springer.
18. Abubakar, A., Ugail, H., & Bukar, A. M. (2019b). Noninvasive assessment and classification of human skin burns using images of Caucasian and African patients. *Journal of Electronic Imaging, 29*(4), 041002.
19. Abubakar, A., Ugail, H., & Bukar, A. M. (2020). Assessment of human skin burns: A deep transfer learning approach. *Journal of Medical and Biological Engineering*. https://doi.org/10.1007/s40846-020-00520-z
20. Abubakar, A., Ajuji, M., & Usman Yahya, I. (2020). Comparison of deep transfer learning techniques in human skin burns discrimination. *Applied System Innovation, 3*(2), 20.
21. Kuan, P., et al. (2017). A comparative study of the classification of skin burn depth in human. *Journal of Telecommunication, Electronic and Computer Engineering, 9*(2–10), 15–23.
22. Yadav, D., et al. (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. Khan, F. A., et al. (2020). Computer-aided diagnosis for burnt skin images using deep convolutional neural network. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-020-08768-y
24. Gill, P. J. (2013). The critical evaluation of laser Doppler imaging in determining burn depth. *International Journal of Burns and Trauma, 3*(2), 72.
25. Gu, J., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition, 77,* 354–377.
26. Bukar, A. M. (2019). *Automatic age progression and estimation from faces*, 2019, University of Bradford.
27. Jilani, S., Ugail, H., & Logan, A. (2019). *The computer nose best*. In: *2019 13th international conference on software, knowledge, information management and applications (SKIMA)*. 2019. IEEE.
28. Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition.* arXiv preprint https://arxiv.org/1409.1556.
29. Deng, J., et al. (2009). *Imagenet: A large-scale hierarchical image database*. in *2009 IEEE conference on computer vision and pattern recognition*. 2009. IEEE.

30. He, K., et al. (2016). *Deep residual learning for image recognition*. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
31. Mahmood, A., et al. (2020). ResFeats: Residual network based features for underwater image classification. *Image and Vision Computing, 93,* 103811.
32. Dindorf, C., et al. (2020). Interpretability of input representations for gait classification in patients after total hip arthroplasty. *Sensors, 20,* 4385.
33. Vapnik, V. (2013). *The nature of statistical learning theory*. New York: Springer.
34. Blanco, V., Japón, A., & Puerto, J. (2018). *Optimal arrangements of hyperplanes for multiclass classification.* arXiv preprint https://arxiv.org/1810.09167.
35. Ragab, D. A., et al. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ, 7,* e6201.
36. Zhang, C., et al. (2020). Received signal strength-based indoor localization using hierarchical classification. *Sensors, 20*(4), 1067.
37. Mahfouz, A. M., Venugopal, D., & Shiva, S. G. (2020). Comparative analysis of ML classifiers for network intrusion detection. *Fourth international congress on information and communication technology*. Berlin: Springer.
38. Alabi, R. O., et al. (2020). Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. *International Journal of Medical Informatics, 136,* 104068.
39. Soleymani, R., Granger, E., & Fumera, G. (2020). F-measure curves: A tool to visualize classifier performance under imbalance. *Pattern Recognition, 100,* 107146.