DFUNet: End-to-End Diabetic Foot Ulcer Segmentation Framework with Vision Transformer Based Detection

No Author Given

No Institute Given

Abstract. Diabetic foot ulceration (DFU) is an open wound that occurs in approximately 15% of patients with diabetes, and is mostly located on the sole. Currently, it is one of the major challenges for healthcare systems around the world and a serious complication of diabetes. DFU causes infection and ischemia which can significantly prolong treatment and result in limb amputation and terminal illness. Thus, regular monitoring of the DFU area is necessary to assess the healing process and improve the care. In this regard, we propose an end-to-end ensemble fully convolutional network (DFUNet), which mainly includes three modules: the U-Net module (DFU_UNet), the hybrid approach (DFU_detect) containing the YOLOv4 based detection module and the Vision Transformer DETR detection approaches. Ensemble solution based on the high ranking strategy is based on combining DFU_UNet with a hybrid solution being a combination bounding-box detection which is performed using the latest DETR vision transformer architecture and YOLOv4 followed by patch segmentation. We achieved 0.643 Dice score for the DFUNet ensemble based approach, 0.648 for DFU_UNet, and 0.556 and 0.581 for hybrid approaches based on YOLOv4 and DETR, respectively. The ensemble DFUNet is not detecting 18 out of 200 changes, while the DFU_UNet 16 changes, the DETR 3, and YOLOv4 42 changes. The results are based on the validation set of the DFUC 2022.

Keywords: DFU · diabetic foot · segmentation · detection · DETR · U-Net · YOLO

1 Introduction and Problem Statement

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. It is a very common disease which affects over 425 million people worldwide [6]. Based on World Health Organization in 2019, diabetes was the direct cause of 1.5 million deaths and 48% of all deaths due to diabetes occurred before the age of 70 years. One of its most severe complications is a diabetic foot - a conditions that can range from minor walking problems to a complete malfunction of nerve system and a disability to feel pain or touch. The risk of a diabetic patient developing a foot ulcer along their lifetime is estimated to be 19-34% [2]. If not treated properly it often results in a limb amputation. Apart from

balancing the diabetes itself, the treatment required careful observation of the affected areas and very special care. This is especially important at early stages. Thus, there is a huge demand for CAD system dedicated for detection, precise segmentation as well as comparison and assessment of DFU areas. Such system can be beneficial for several reasons: a) visual inspection is always subjective and inaccurate for DFU area segmentation, assessment and tissue classification, b) DFU examination observations are not always recorded in a consistent format and thus no comparison possible, and c) manual assessment of a patient's DFU areas during the healing period is difficult without computer support.

Our contribution to this research area can be summarized as:(1) we propose an ensemble end-to-end DFU segmentation framework which includes three modules: the U-Net module (DFU_UNet), the hybrid approach (DFU_detect) containing the YOLOv4 based detection module and the Vision Transformer DETR detection approaches, (2) we confirm that Vision Transformers can achieve high results for detection tasks in the medical domain, (3) We compare the outcomes of state-of-the-art models including end-to-end U-Net architecture, detection with U-Net segmentation for YOLOv4 and DETR Vision Transformer for the DFU area segmentation task. We visualize the segmentation areas extracted by each architecture, (4) Our solution scored 0.643 Dice score for the DFUNet ensemble based approach, 0.648 for DFU_UNet, and 0.556 and 0.581 for hybrid approaches based on YOLOv4 and DETR, respectively.

1.1 Related works

Due to the contribution to this research area in a form of a Diabetic Foot Ulcer Challenge which has been organised since 2020 many research groups have addressed the DFU area assessment including detection (2020) [5] and classification (2021) [24]. Here, we present the most important research papers both regarding DFU Challenge including classification and detection as well as general segmentation problem.

In 2021 the highest classification result 0.6216 F1-Score and 0.8855 AUC has been achieved by [9]. The author also used Vision Transformers, but they were outperformed by Convolutional Neural Networks (CNN). Very interesting results and AI architectures with unique approaches have been presented by [3, 1, 20, 12]. In [3] Authors used an ensemble of EfficientNet architecture [23] and pix2pix model [15]. In [20], the Authors tested several pre-trained Vision Transformers and fine-tuned them for the classification task. It is, however, worth mentioning, that to the best of our knowledge, there were no solutions participating in the challenge that used Vision transformers for detection or segmentation.

For the segmentation task in general, there are many different solutions, ranging from classical computer vision techniques to sophisticated deep learning architectures including, CNN and U-Net based models or Vision Transformers. The most popular architectures used for medical image segmentation are U-Nets [22] and autoencoders, with many different variations such as nnU-Net [14], CE-Net [11], UNET 3+ [13], or the latest Half-UNet [19]. Detection task, on the other hand is mostly dominated by RCNN-like methods, recently outperformed

by faster solutions like YOLOv4-7 or SSD [18, 21] as well as with the Vision Transformers [8]. Although not so widely used in the segmentation task itself, they already begin to outperform traditional convolutional networks in tasks such as classification and detection. In section 2.3 we describe how this new solution, especially focusing on DETR architecture [4], might be utilised for DFU detection and segmentation stage.

2 DFUNet: End-to-End Ensemble DFU Segmentation Framework

DFU segmentation task vary from segmenting small areas to big ones that cover most of the entire image. With such a diverse dataset it can be difficult to tune the network to perform well. Especially, there is a risk of false positive, as the network may focus on bigger, more definitive objects in the background. Therefore, as a solution to this size diversity we propose the ensemble learning methodology. The ensemble learning methods are used to improve the results of deep learning by combining several models, which gives an improved prediction. Thus, the proposed DFUNet detection and segmentation framework uses ensemble segmentation technique, which improves the accuracy of segmentation with enhanced performance. The proposed DFUNet framework consists of three modules presented in Fig. 2: the end-to-end U-Net module (DFU_UNet), the hybrid approach (DFU_detect) containing the YOLOv4 based detection module and the Vision Transformer DETR detection approach. Similar to the ensemble, the final prediction is the aggregation of the prediction made by each three proposed modules including the high ranking strategy based on combining DFU_UNet with a hybrid solution being a combination bounding-box detection which is performed using the latest DETR ViT architecture and YOLOv4 followed by patch segmentation. The iterative architecture adjustment, ensemble learning, and hyperparameter tuning have been used in this model, which has resulted to improve the segmentation outcome.

2.1 Dataset: Diabetic Foot Ulcer Segmentation Challenge 2022

Diabetic Foot Ulcer Segmentation Challenge 2022 dataset consists of a training set of 2000 RGB images and a testing set of 2000 RGB images, where ulcer regions were delineated by experienced podiatrists and consultants. The DFUC2022 training set consists of 2304 ulcers, where the smallest ulcer size is 0.04% of the total image size, and the largest ulcer size is 35.04% of the total image size. The ratio of the ulcer region to the total image size is following: 89% (2054 out of 2304) of the ulcers are less than 5% of the total image size. The smaller images in particular represent a significant challenge for segmentation algorithms as it is widely known that deep learning algorithms tend to omit small regions [10]. A small number of duplicates, which were annotated by different experts, were also added to the dataset [16]. More details on the DFUC organization, as well as live test leaderboard, can be found online at https://dfu-2022.grand-challenge.org/ .

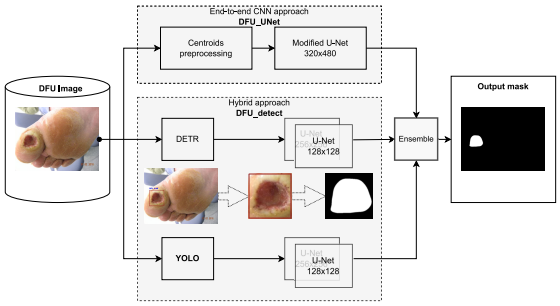


Fig. 1. DFUNet architecture consisting of three main modules: the U-Net part (DFU_UNet) and the hybrid approach (DFU_detect) containing the YOLOv4 based detection module and Vision Transformer DETR detection module. Before output each of them are combined using ensemble methods.

2.2 Module I: End-to-end segmentation framework based on U-Net architecture

The U-Net based architecture introduced in 2015 by [22] with many novel improvements is still today the most widely used in biomedical image segmentation tasks outperforming other solutions.

Data preprocessing for segmentation: As the models accepts inputs smaller than dataset original image size, we take advantage and provide diversified training set for each epoch, simply by extracting its subset. Firstly, we use original mask in order to find location of a reference object. Then, having its centroid and bounding box coordinates we randomly choose a patch around that centroid which serves as the input to the network. Randomness guarantees that for each epoch, for particular image, an extracted example differs from a previous one from the same image to the previous epoch. Such approach, together with typical augmentation methods like flips, re-scaling and translation, helped us to reduce overfitting during training. In order to provide the most representative examples during training, we divided patches extraction into three alternatives (Fig. 2). The first one randomly picks a new centroid which is inside the bounding box of the object. This guarantee that small objects should stay near the center of the patch, but never in the exactly same place. The second one widens the interval where a new centroid is chosen from to the size of the patch. Thanks to that, objects may appear on the edge of the patch. The last possibility randomly choose a new centroid from the area of the whole image in order to provide examples with no objects on them. Such a new centroid is then used to extract a patch of neural network input size. During our experiments we achieved the best results for using those steps with ratio of 0.75:0.15:0.1 which means that in each batch there where, on average, 75% examples extracted using first approach, 15% with the second one and 10% with the third. After a single patch is extracted, it is converted to a CIELAB color space and then concatenate with the RGB rep-

resentation. Finally, a 6-channels patch extracted from an image constitute a training example to the network.

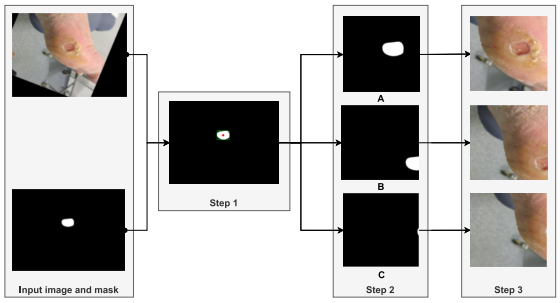


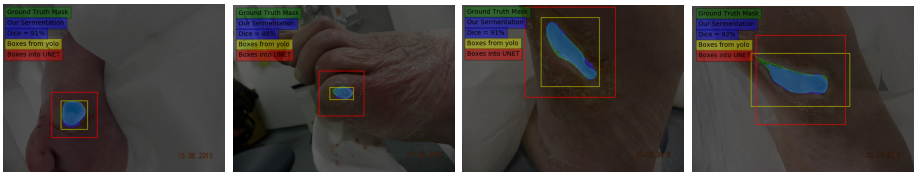
Fig. 2. Image preprocessing stage: Input image and mask augmentation, Step 1: centroid and bounding box coordinates of the object on the ground-truth mask, Step 2: randomly choose next step (2a, 2b or 2c) where 2a: centroid coordinates are drawn within the object's bounding box, 2b: alternative centroid is picked within range between the original centroid and patch size, 2c: new centroid randomly chosen on the whole image. Step 3: extracting the patch which constitutes an input to a network.

Training and evaluation: In the end-to-end approach we took advantage of a U-Net and Xception hybrid architecture [7]. The input was defined as 320 *times* 480 px based on our experiments. Smaller network input, than the original images, was used on purpose in order to use preprocessing technique described above, mainly for the possibility of randomly choosing different patches from single image. For training a 5-fold validation approach was used. The dataset was divided into 5 subsets, each of which consisted of 1600 examples for training and 400 for validation. Then five models were trained for 860 epochs, starting with $10e^{-3}$ learning rate and lowering it to $10e^{-5}$ after 130 epochs and then to $10e^{-5}$ after 650 epochs. As a cost function a combination of binary crossentropy and Jaccard Loss was proposed. Jaccard coefficient is defined as: $J(A, B) = |A \cup B|/|A \cap B|$, where A and B are reference set and modelś output set respectively. As Jaccard coefficient maximum value is 1.0 it can be used as loss function as: $J_{loss} = 1 - J$. Because Jaccard works only for predictions which already have intersection with reference (for those without, it equals to 0 so there is no gradient to calculate weights update) a Binary Crossentropy loss (BCE_{loss}) was added. The final loss function was defined as: $Loss = 0.2 * BCE_{loss} + 0.8 * J_{loss}$.

For evaluation purposes, the challenge's validation set was used. Because the input was smaller than the original images we used sliding window of model's input size and evaluated each extracted patch with stride of 16 $\times$ 16 px in both height and width. Overlapping results was then averaged and rounded to 1 if above 0.5. Such a result constituted a segmentation mask for a single model and achieved 0.6478 Dice score on challenges validation set.

2.3 Hybrid solution combining detection and patch segmentation with YOLOv4 and Vision Transformers

Based on the size diversity of the wound areas we propose a hybrid model consisting of detection stage (YOLOv4 and DETR Vision Transformer architecture) and segmentation based on U-Net architecture, albeit with a few modifications.



a) b)

Fig. 3. Examples presenting the BBoxes obtained with YOLOv4 and segmentation results: a) small detections resized into patches of size 128 × 128, b) large detections resized into patches of size 256 × 256.

Module 2: YOLOv4 based detection: The YOLO [21] approach involves a single neural network trained end-to-end that takes an image as input and predicts bounding boxes and class labels for each bounding box directly. The network was originally designed as a very fast solution, dedicated to real-time applications such as autonomous vehicles. Since then, it has become the most commonly used benchmark solution for almost all detection tasks. To improve and shorten the learning process the model was based on YOLOv4 model pretrained on COCO dataset [17]. Most importantly, instead of training one model for patch segmentation we tested to different patch sizes including 128 × 128 for smaller detections and 256 ×256 for larger ones. In Fig. 3 we present the outcomes for detection process and patch generation.

Module 3: DETR Vision Transformer based detection: DETR was introduced in 2020 by Facebook AI in the paper [4]. The framework consists of CNN backbone network, positional encoding module, encoder-decoder architecture based on transformers and prediction heads. The pipeline utilizes biparite matching between the predicted and ground-truth objects in order to drop post-processing of detections. Due to its parallel nature of processing predictions, DETR is very efficient and fast.

In order to increase accuracy, we decided to utilize DETR model pretrained on COCO dataset [17], which we fine-tuned on challenge data. We used the original loss with null class coefficient set to 0.05. We have left default setting of $num\_queries = 100$. Even though the detection results obtained with DETR are fairly accurate, with our choice of parameters the number of false positives and duplicated detections is large. In order to drop unnecessary bounding boxes,

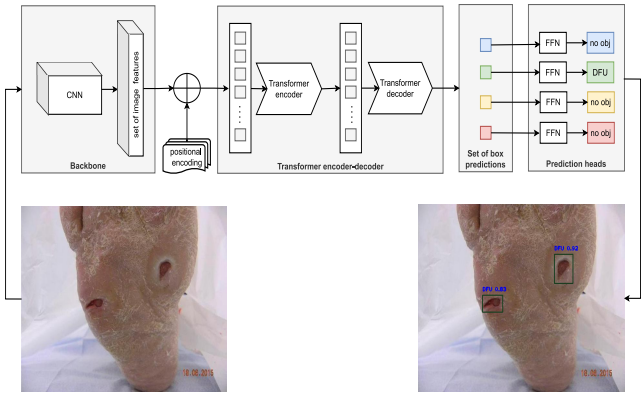


Fig. 4. DETR Architecture containing the most essential segments: CNN ResNet-50 backbone, Transformer encoder-decoder and prediction heads with FFN (feed forward network), which is a 3-layer perceptron with ReLU activation function.

instead of decreasing number of queries, we consider only the 3% of highest confidence detections. This setting improves the detection of small ulcers and the ones barely distinguishable from the skin, and also enables filtering of the duplicate and false-positive predictions. In contrast to YOLO which requires anchors, DETR predicts all objects at once, and is trained end-to-end with a set loss function which performs bipartite matching between predicted and ground-truth objects. This can be a reason why DETR model achieved far better precision, detecting images that were ignored by YOLOv4 module (see Fig. 5).

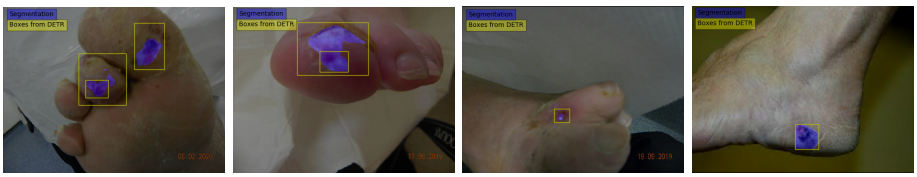


Fig. 5. Ulcer examples from the official validation set that were missed by YOLOv4 and detected by DETR.

Patch adjustment and egmentation: To address the problem of very different ulcer areas, we have decided to use two sizes of patches: 128 × 128 for smaller objects and another of size 256 × 256 for larger ones. The segmentation architecture is based on U-Net as described in section 2.2. Both models were trained using 5-fold cross-validation, and later verified on a DFU 2022 validation set. Training hyperparameters were optimised - the model was trained for 50 epochs, with Adam optimizer and batch size of 8. Finally, the predictions

of the 5-fold cross-validated model were thresholded to address the problem of undersegmentation. The confidence level was lowered up to 0.2.

3 Experimental Results and Analysis

In this section we describe conducted experiments and present calculated statistics. Metrics such as Mean Overlap, Union Overlap, Dice Coefficient, Volume Similarity, False Negative Error, False Positive Error, and Jaccard Coefficient are calculated by uploading the mask results on the challenge website, so they can be easily compared with other solutions taking part in the challenge.

In Table 1 we present achieved results for the validation set for 3 modules: YOLOv4, DETR and DFU_UNet as well as ensemble outcome for DFUNet. The DFU_UNet architecture achieved similar results to the ensemble method scoring 0.64 Dice score.

Table 1. Comparison of results obtained on the validation set from challenge website

				DFUNET: Ensemble	
Metric	**YOLOv4**	**DETR**	**DFU_UNet**	**Sum**	**Vote**
MeanOverlap	0.5556	0.5808	0.6479	0.6140	0.6432
UnionOverlap	0.4746	0.4790	0.5483	0.5090	0.5471
DiceCoefficient	0.5556	0.5808	0.6479	0.6140	0.6433
VolumeSimilarity	0.2473	0.0139	0.0898	-0.1757	0.1208
FalseNegativeError	0.3975	0.3622	0.2991	0.3873	0.2952
FalsePositiveError	0.3931	0.3293	0.2945	0.2379	0.3020
JaccardCoefficient	0.4746	0.4790	0.5483	0.5090	0.5471

However, we observed, that the DFU_UNet tend to omit some DFU areas while the DETR ViT architecture finds even the smallest changes in the skin, despite a very high margin of confidence threshold (0.97).

4 Conclusion

In this paper, we proposed the DFUNet architecture which consists of three segmentation approaches which can effectively detect and extract DFU areas. The segmentation results highlights the difficulty in DFU clinical delineation and assessment. This work sheds light on the challenges inherent in the development of AI systems which can aid the standardisation of DFU delineation over time to track healing progress. Next, we will analyze in-depth the obtained results, propose data augmentation for omitted cases and explore deep semi-supervised learning methods to enhance detection and segmentation accuracy.

References

1. Ahmed, S., Naveed, H.: Bias adjustable activation network for imbalanced data - diabetic foot ulcer challenge 2021. In: DFUC@MICCAI (2021)
2. Armstrong, D.G., Boulton, A.J.M., Bus, S.A.: Diabetic foot ulcers and their recurrence. The New England journal of medicine **376 24**, 2367–2375 (2017)
3. Bloch, L., Brungel, R., Friedrich, C.: Boosting efficientnets ensemble performance via pseudo-labels and synthetic images by pix2pixhd for infection and ischaemia classification in diabetic foot ulcers. In: DFUC@MICCAI (2021)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. ArXiv **abs/2005.12872** (2020)
5. Cassidy, B., Reeves, N.D., Pappachan, J.M., Gillespie, D., O'Shea, C., Rajbhandari, S., Maiya, A.G., Frank, E., Boulton, A.J.M., Armstrong, D., Najafi, B., Wu, J., Kochhar, R.S., Yap, M.H.: The dfuc 2020 dataset: Analysis towards diabetic foot ulcer detection. TouchREVIEWS in endocrinology **17 1**, 5–11 (2021)
6. Cho, N.H., Shaw, J.E., Karuranga, S., Huang, Y., Fernandes, J.D.D.R., Ohlrogge, A.W., Malanda, B.: Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes research and clinical practice **138**, 271–281 (2018)
7. Chollet, F., et al.: Keras (2015), https://github.com/fchollet/keras
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2021)
9. Galdran, A., Carneiro, G., Ballester, M.Á.G.: Convolutional nets versus vision transformers for diabetic foot ulcer classification. In: DFUC@MICCAI (2021)
10. Goyal, M., Yap, M.H.: Multi-class semantic segmentation of skin lesions via fully convolutional networks. ArXiv **abs/1711.10449** (2020)
11. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE Transactions on Medical Imaging **38**, 2281–2292 (2019)
12. Güley, O., Pati, S., Bakas, S.: Classification of infection and ischemia in diabetic foot ulcers using vgg architectures. Diabetic foot ulcers grand challenge : second challenge, DFUC 2021, held in conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021 : proceedings. DFUC (Conference) **13183**, 76–89 (2021)
13. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 1055–1059 (2020)
14. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods (2020)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5967–5976 (2017)
16. Kendrick, C., Cassidy, B., Pappachan, J.M., O'Shea, C., Fernández, C.J., Chacko, E.C., Jacob, K., Reeves, N.D., Yap, M.H.: Translating clinical delineation of diabetic foot ulcers into machine interpretable segmentation. ArXiv **abs/2204.11618** (2022)

17. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
18. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV (2016)
19. Lu, H., She, Y., Tie, J., Xu, S.: Half-unet: A simplified u-net architecture for medical image segmentation. Frontiers in Neuroinformatics **16** (2022)
20. Qayyum, A., Benzinou, A., Mazher, M., Mériaudeau, F.: Efficient multi-model vision transformer based on feature fusion for classification of dfuc2021 challenge. In: DFUC@MICCAI (2021)
21. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 779–788 (2016)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. ArXiv **abs/1505.04597** (2015)
23. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. ArXiv **abs/1905.11946** (2019)
24. Yap, M.H., Cassidy, B., Pappachan, J.M., O'Shea, C., Gillespie, D., Reeves, N.D.: Analysis towards classification of infection and ischaemia of diabetic foot ulcers. 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) pp. 1–4 (2021)