 doi:10.20944/preprints202101.0346.v1

Article

A Highly Transparent and Explainable Artificial Intelligence Tool for Chronic Wound Classification: XAI-CWC

Salih Sarp [1], Murat Kuzlu [1], Emmanuel Wilson [2] , Umit Cali[3,*] and Ozgur Guler [2]

[1] Batten College of Engineering & Technology, Old Dominion University Norfolk, VA, USA
[2] eKare, Inc Fairfax, VA, USA
[3] Norwegian University of Science and Technology,Trondheim, Norway
* Correspondence: umit.cali@ntnu.no

Abstract: Artificial Intelligence (AI) has seen increased application and widespread adoption over the past decade despite, at times, offering a limited understanding of its inner working. AI algorithms are, in large part, built on weights, and these weights are calculated as a result of large matrix multiplications. Computationally intensive processes are typically harder to interpret. Explainable Artificial Intelligence (XAI) aims to solve this black box approach through the use of various techniques and tools. In this study, XAI techniques are applied to chronic wound classification. The proposed model classifies chronic wounds through the use of transfer learning and fully connected layers. Classified chronic wound images serve as input to the XAI model for an explanation. Interpretable results can help shed new perspectives to clinicians during the diagnostic phase. The proposed method successfully provides chronic wound classification and its associated explanation. This hybrid approach is shown to aid with the interpretation and understanding of AI decision-making processes.

Keywords: Chronic wound classification; transfer learning; explainable artificial intelligence.

1. Introduction

Alan Turing introduced many central concepts of Artificial Intelligence (AI) in his "Computing Machinery and Intelligence" article (1950) [1]. After alternating between periods of great passion and setback [2], AI has found its place as a critical component of growth in a variety of applications [3]. These applications range from diagnostic decision assistants in healthcare, safety-critical systems in autonomous vehicles, and long-term financial investment planning, benefit from these breakthroughs [4].

AI is capable of analyzing complex data and exploiting non-intuitive approaches to derive meaningful relationships [5]. Healthcare applications based on AI are utilized in early detection, diagnosis, treatment, as well as outcome prediction and prognosis evaluation [6]. The barrier stands the way for AI applications sourced from the lack of transparency and black-box nature that cannot be explained directly [7]. When an AI model learns and gives an output, it processes the data and deciphers the processed information immediately, instead of storing the learned data as a clear digital memory [8]. This is why an explainable and understandable glass-box approach should be taken to enable transparent, trustable, and re-traceable AI applications [9].

The Explainable Artificial Intelligence (XAI) term is coined to provide transparency and guided inference in understanding the decision-making processes of the AI system [10]. The study in [11] provided a comprehensive review of XAI in terms of concepts, taxonomies, opportunities, and challenges, as well as discussed adopting XAI techniques to image processing. The study [12] summarized the recent developments in XAI and its connection with artificial general intelligence, as well as identified trust-related problems of AI applications. Authors in [13] identified nuances, challenges, and requirements for the design of interpretable and explainable machine learning models and systems in

 doi:10.20944/preprints202101.0346.v1

healthcare and described how to choose the right interpretable machine learning algorithm. Compared to conventional black-box AI systems, the XAI system is a glass box providing data about the intermediate steps of the inference process [14]. An example of this would be a computer-aided diagnosis system that not only outputs a prediction but also shows where it looked during the decision-making process by overlaying a heat map on top of an X-ray image. The study in [15] presents the Grad-CAM technique by utilizing the gradients that are taken from the convolution layer to generate a highlighted localization map. Grad-CAM benefits the convolutions, whereas our proposed method calculates the most effective features by tweaking the input and perceiving its effect on classification. Authors in [16] presented classification tasks using LIME (Local Interpretable Model-Agnostic Explanations) to explain predictions of Deep Learning (DL) models, to be able to make these complex models partly understandable.

In [17], the authors proposed a classification technique where they combined the Genetic Algorithm (GA) and Adaptive Neural Fuzzy Inference System (ANFIS) to predict heart attack through XAI at quite satisfactory rates. Authors in [18] developed an assisted and incremental medical diagnosis system using XAI, which allows interaction between the physician (i.e., human agent), and the AI agent. Authors in [19] investigated the problem of explainability in AI in the medical domain where wrong decisions of the system can be very harmful and proposed two approaches to explain predictions of deep learning models, (i) computes sensitivity of the prediction with respect to changes in input, and (ii) decomposes decision in terms of the input variables. Authors in [20] investigated how to increase the trust in computer vision through XAI, and how to implement XAI to better understand AI in a critical area such as disease detection.

This paper presents a highly transparent and explainable artificial intelligence tool for chronic wound classification. Objectives of the study are:

- Build a wound type classification model using deep learning and transfer learning methods.
- Apply XAI methods to convert complex black-box AI systems to more understandable glass box AI systems.
- Provide insights into the complex decision-making processes of an AI system.

2. Methodology

This section discusses the methodology of transfer learning for the wound type classification and XAI for providing transparency to the classification task.

2.1. Transfer Learning

Predictions on new data utilizing data distributions and statistical properties of a previously trained model are called transfer learning [21]. Studies over transfer learning have been increasing because of the lack of limited computation power and limited dataset size. The same distribution of the training and the testing dataset is needed for traditional machine learning models [22]. However, transfer learning provides flexibility and capability of training on a smaller dataset by transfer of learned features from an old model to the new model.

Transfer learning could be leveraged at three levels of training. Learning performance can be improved dramatically at the very beginning of training. Secondly, a steeper learning curve can be achieved using transfer learning. Thirdly, the final step performance of the model training is improved significantly by transfer learning as well [23].

Although transfer learning is beneficial, the abundance of labeled data in the source domain is used to improve the shortage of labeled data in the target domain. There are two things that we faced during the application of transfer learning. These are negative transfer learning and imbalanced distribution of classes [24]. Negative transfer learning sourced from the irrelevance of the translated features to the target domain. The class imbalance happens when there is a bias towards a class in classification tasks. Both scenarios decrease the performance of the newly built model.

 doi:10.20944/preprints202101.0346.v1

The transfer learning application comprises two steps, (i) feature extraction and (ii) fine-tuning. The pre-trained network will extract meaningful features from the new data samples, and a final classifier is added on top of the pre-trained network to do classification tasks in the target domain. The pre-trained network already mastered the feature extraction task with convolutional networks. The second step is fine-tuning with freezing and unfreezing some of the top layers from the pre-trained model to train for higher performance jointly. ResNet, EfficientNet, and VGG16 (Very Deep Convolutional Neural Networks for Large-Scale Image Recognition) networks are a few of the successful DL models for classification tasks.

In this study, VGG16 is used, as shown in Figure 1, which gives the flexibility and best score among other DL models. VGG16 consists of roughly 138 million parameters and is trained over 14 million images on the ImageNet database. The pre-trained convolution layers of the VGG16 architecture are kept frozen, and only fully connected layers are trained in the first phase of the transfer learning where convolution layers' weights are not updated. In the second phase, the convolution layers are kept frozen, but the last convolutional layer. The last convolution layer and fully connected layers are trained together for fine-tuning the model. The weights of convolution layers from the VGG16 are transferred to benefit from their feature extraction skills. The training of the last convolution layer provides the fine-tuning to obtain better classification results.

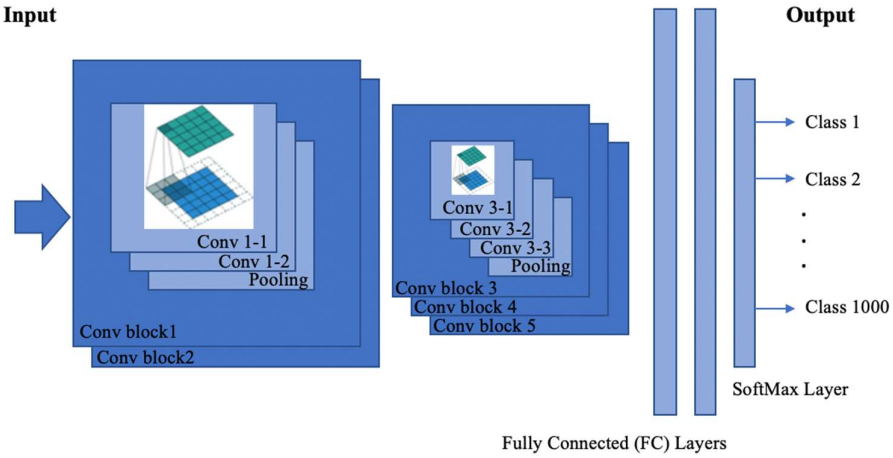


Figure 1. *VGG16 architecture.*

2.2. Explainable AI

Artificial Intelligence (AI) provides tremendous benefits in various sectors, but its adoption is limited due to the non-intuitive, opaque nature of machine learning models [25]. The internal working of an AI model is complicated and requires a strong mathematical background to understand. This can be a significant barrier to entry [26]. There are two kinds of approaches to explain an AI model; (i) the comprehensible, and (ii) the interpretable model. Comprehensible models are explained with posthoc explainability approaches. Classical machine learning methods (e.g., regression models and decision trees), are interpretable models, as these reveal more transparency compared to convolutional networks [27]. The inner workings of the machine learning models might be complicated and hard to interpret, but its efficiency and accuracy are higher than human performance in many cases [28]. That's why we need to comprehend the inner workings of the machine learning models, even if the model itself is not interpretable.

Generalized Linear Models (GLM) provide meaningful, clear, and accessible feature importance that indicates the relative importance of each feature when making a prediction for the regression models. Outputs of regression models are a linear combination of features with different weights depending on the significance of features itself [29].

 doi:10.20944/preprints202101.0346.v1

Tree-based models have individually meaningful features, and tabular-style datasets are used in these models. The connection of tree-based models to the training data results in more interpretability with local explanations in comparison to linear regression models [30].

DL models exceed human performance, where they have an enormous number of parameters and nested non-linear structures. DL is a relatively new research field compared to classical machine learning models. Outputs of a deep learning model are sometimes unexpected and counter-intuitive to human reasoning (e.g., the case in a Go game between Lee Sedol and AlphaGo, AI player) [31]. The decision-making processes of AI may be unanticipated, but it could provide insights and improve how we handle the tasks from a bottom-up approach.

The sheer number of parameters and non-linear structure of deep learning prevent linking inputs to the model prediction. Therefore, a post hoc explainability approach is taken. Gradient and attention-based methods are developed and used in the context of image and text-based models, respectively. The gradient-based method brings attention to important regions in the input image in the backward pass. Attention-based method train attention weights which determines how much each of the elements are in the final output [32].

Generalized explainable AI methods are designed to treat any machine learning model as a black-box with inputs and some outputs [33]. One of these methods is Local Interpretable Model-Agnostic Explanations (LIME) [34]. LIME finds the statistical connection between the input and the model prediction by training local surrogate models on perturbed inputs instead of training them globally [35]. It provides both an explanation of an instance by an interpretable representation and visualization. This study provides the explainability and transparency of chronic wound classification using transfer learning implementation with Keras and XAI methods.

3. Data Collection, Pre-processing, Environment, and Validation

This section discusses data collection, data pre-processing, and the test environment. The details about the dataset are given in the data collection section. Forming a ground truth for classification and the environment that the model runs on is explained in data pre-processing and environment sections accordingly.

3.1. Data Collection

The chronic wound data repository, which includes diabetic, lymphovascular, pressure injury, and surgical wound types are collected from the data repository of eKare, Inc. and anonymized for patient privacy [36]. eKare Inc. has specialized in wound management, and its services are in use by many hospitals where the wound images are collected and processed. 8690 wound images are chosen by an MD specialized in wound care to represent the mentioned wound types. There are 1811 diabetic, 2934 lymphovascular, 2299 pressure injury, 1646 surgical wound images in the dataset used in this study. The proposed model uses wound images to predict wound etiology utilizing transfer learning, data augmentation, and deep neural networks (DNN).

3.2. Data Pre-processing

The dataset was reviewed by a trained MD to ensure correct classification of underlying chronic wound etiology. Classification of the wounds is carried out to form the clinical ground truth for the wound images. Wound images are then hand-labeled for the wound type classification.

The distribution of the dataset is not even as the dataset is fine-tuned for a correct representation of chronic wound classes. Data augmentation techniques are used to increase the dataset size and maintain class balance. The dataset, 8692 images in total, was split into training and test sets comprising 6520 and 2172 images, respectively. Then the collected data was pre-processed to increase data quality. It includes formatting, rescaling,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 January 2021 doi:10.20944/preprints202101.0346.v1

and normalization of the images. Images have been scaled to 224x224 pixels and normalized for a faster learning process.

3.1. Environment

XAI-CWC model was implemented using Keras deep learning framework with Python version 3.6. We used a workstation to run our model, which has an Intel® Core ™ i7 -8700X CPU @3.20 GHz with 32 GB memory and NVIDIA GeForce GTX 1080 GPU with 8 GB dedicated and 16 GB shared memory. We trained the model for 1000 epochs where the model has warmed up 250 epochs with only training fully connected (FC) layers, then an additional 750 epoch with the training of FC layers and the final set of the convolutional layers. The total training of the model took around 8 hours. We used a constant learning rate of 0.001 for the "RMSprop" optimizer for the training.

4. Implementation of Transfer Learning And XAI Approaches On Wound Classification

The objective of this paper is to explore and apply XAI methods on chronic wound classification to expand knowledge about the opaque, "black-box" structure of the machine learning models. The finely chosen dataset is split into two primary sets, i.e., training and test set. The test dataset comprised 25% of the data, while the remaining 75% was used as training data. Data augmentation techniques, such as mirroring, rotation, and horizontal flip, are used to avoid overfitting and increase the dataset for better training performance. Test data is indexed for generalization of the model and proper comparison. Transfer learning is realized in two steps, first, a warm-up phase, and second, a fine-tuning phase. This study using transfer learning provided satisfying results according to performance metrics, which are F1 score, recall, and precision, which are extracted from the confusion matrix shown in Table 1.

Table 1. *Confusion Matrix*

		Prediction	
		y'=0	y'=1
True label	y=0	True Negative	False Positive
	y=1	False Negative	True Positive

Performance measures are given in Eq. 1-3 below.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (1)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (2)$$

$$\text{F1 score} = 2\text{x}\frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Precision, Recall, and F1-scores of the proposed model are given in Table 2.

Table 2. *Performance evaluation of CWC-XAI model*

Model	**Precision**	**Recall**	**F1-score**
Diabetic	0.72	0.71	0.71
Lymphovascular	0.87	0.82	0.84
Pressure Injury	0.96	0.63	0.76
Surgical	0.77	0.70	0.73
Average	**0.83**	**0.71**	**0.76**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 January 2021 doi:10.20944/preprints202101.0346.v1

Precision gives the ratio of correctly classified wound types over total positive wound type predictions. The recall is the measure of how many of the positive wounds are correctly classified. This metric checks predictions in the eye of true labels. If the recall is high, true positives are identified well by the model, and there are fewer wrongly classified samples. Both of these metrics could be high, yet the model could still underperform. This is why a third metric is utilized to measure model performance. F1-score is a hybrid measurement that brings together both precision and recalls for a better evaluation.

Precision, recall, and F1-scores of each wound type, and their averages, are compared in Table 2. Higher precision values of lymphovascular and pressure injury wound types indicate that the model performed very well with these wound types, whereas pressure injuries are harder to diagnose as the recall score is high for pressure injury wounds. This means that some of the pressure injury wounds are not learned or are similar to another wound type and misclassified by the model. Lymphovascular wounds have the highest recall among other wound types, which reveals that our proposed method is capable of diagnosing lymphovascular wounds. The F1 score on the performance of lymphovascular wounds is high, and pressure injury is low. Surgical wounds have fair precision and recall scores, but one-third of the wounds are misclassified. Many diabetic wounds stop healing, and surgical intervention is needed. Hence our model is likely to classify a surgical wound as diabetic. The recall of diabetic wound types is pretty high, and it has the lowest F1 score, which is a result of low precision.

AI applications require acceptance during the decision-making from the user beforehand. As results provide efficiency and automation, AI becomes very popular in low-risk fields, such as agriculture, customer services, and manufacturing. However, high-risk applications of AI remain limited, such as health care, as trust is critical in medical practice. Understanding the rationale behind model predictions would certainly help users decide when to trust or not to trust their predictions. The process of explaining individual predictions is illustrated in Figure 2.

A deep neural network using the transfer learning technique was trained using chronic wound images to predict wound type. Accurate wound type designation helps a clinician to classify the wound type, which serves to better steer the treatment approach. Prediction of the image classification is then explained by an "*explainer*" that points to visual features of the image that are the most important to the model. With this information, related to model rationale, the clinician can decide to trust the model or not. Model outputs include an understandable qualitative link between inputs and predictions, which is an essential part of the explainability [37]. The numerous amounts of features in the model are hard to understand by a user, but human reasoning could conclude substantially via a qualitative approach [38]. Another significant property that a reliable explainer should have is local faithfulness. The local explanation is achieved by characterizing the response of a local function with a range of adjacent input [39].

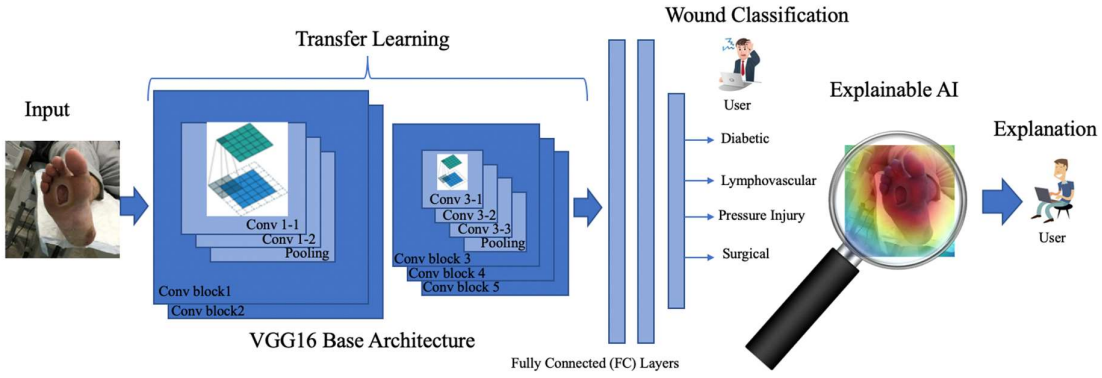
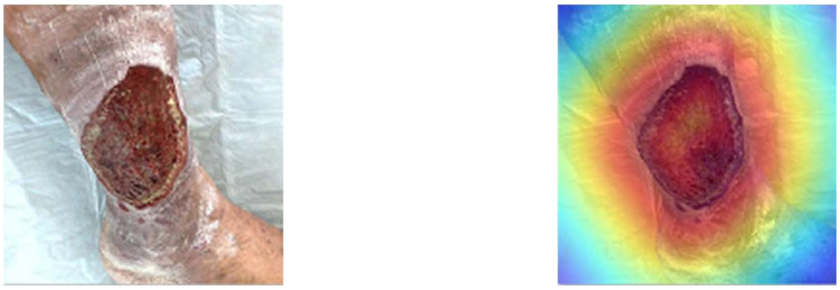


Figure 2. *Wound classification model with DNN, transfer learning, and explainable AI tool.*

 doi:10.20944/preprints202101.0346.v1

In this study, the DNN model with transfer learning and extended XAI technique is used to provide the explainability and transparency for wound image classifiers and why the model would think that a particular class may be present. CWC-XAI model forms a hybrid XAI framework with the use of LIME and heatmap proposal. LIME provides a set of correlated and connected pixels, which is used as an input for the heatmap method. The proposed model provides a focus on the classification task with a heatmap. Warmer colors indicate the more critical areas of the wound in the importance map.

The proposed model classifies a chronic wound as a lymphovascular wound with a probability of 99.9% in Figure 3. It highlights the model's focused area for classification tasks in the wound image with an importance map as an explanation.

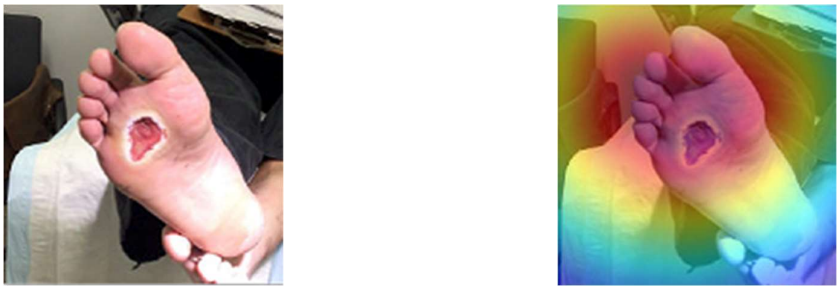


(a) Original wound image **(b)** Importance map

Figure 3. *Explanation of a lymphovascular wound***.**

Figures 4-7 show images of diabetic, lymphovascular, pressure injury, and surgical wounds. Each wound type has a respective heatmap that highlights the focused area that affects the model to choose the proper wound type. Diabetic wound is correctly predicted at 95.36% (Pressure injury: 4.07%, lymphovascular: 0.01%, surgical: 0.55%) and lymphovascular wound is predicted at 100% (Diabetic: 0%, pressure injury: 0%, surgical: 0%) in Figures 4 and 5, respectively. The low diabetic wound classification probability could be increased with additional data that will amplify the feature extraction of diabetic wounds in the training.

The probabilities of wound classification are very high for Figure 6, i.e., pressure injury wound at 100% (Lymphovascular: 0%, surgical: 0%, diabetic: 0%), and for Figure 7, i.e., surgical wound at 99.2% (Diabetic: 0.05%, pressure injury: 0.03%, lymphovascular: 0.01%).



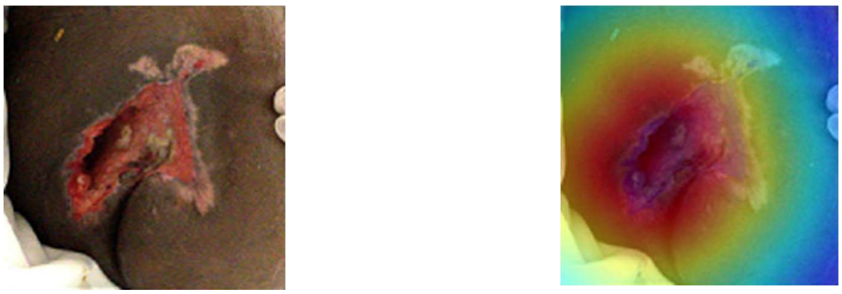
(a) Original Image **(b)** Importance map

Figure 4*. The probabilities of wound types: Diabetic: 95.36%, pressure injury: 4.07%, lymphovascular: 0.01%, surgical: 0.55% (Diabetic).*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 January 2021 doi:10.20944/preprints202101.0346.v1

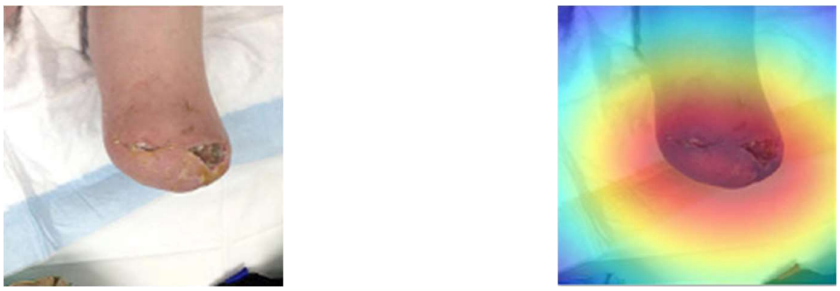
(a) Original Image **(b)** Importance map

Figure 5. *The probabilities of wound types: Lymphovascular: 100%, diabetic: 0%, pressure Injury: 0%, surgical: 0% (Lymphovascular).*



(a) Original Image **(b)** Importance map

Figure 6. *The probabilities of wound types: Pressure Injury: 100%, lymphovascular: 0%, surgical: 0%, diabetic: 0% (Pressure Injury).*



(a) Original Image **(b)** Importance map

Figure 7. *The probabilities of wound types: Surgical: 99.92%, diabetic: 0.05%, pressure injury: 0.03%, lymphovascular: 0.01% (Surgical).*

Figures 4a and 4b show explanations of the most important features that contribute to the prediction. Like Figure 4a/b, Figure 5a/b show explanations and map features with the highest contribution to the prediction for lymphovascular classification. Both figures provide insights as to why the wound type was predicted to be diabetic or lymphovascular. The focus on the diabetic wound includes the wound tissues and toes. The shape of the ulcer and its proximity to toes are the explanations of the diabetic foot ulcer.

The lymphovascular wound, as seen in Figure 4a, is explained with a focus on deeper damaged tissue. This kind of explanation enhances trust in the wound classifier and helps caregivers make a decision and support their decision with a visual explanation. The model using visual cues can explain why a specific wound type is identified as such.

The pressure injury wound explainer focuses on the wounded area and indicates the correct placement of the wound, shown in Figure 6b. In Figure 7, a surgical wound image is explained with a scar pattern and the shape of the wound. The explainer identifies the scar of the wound as the highest feature, and the wound area is highlighted by the proposed model with an importance map.

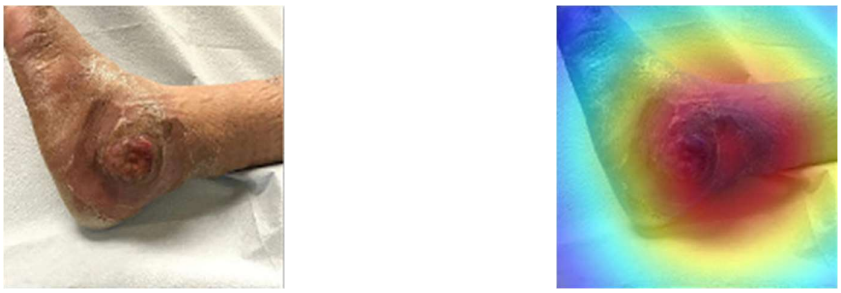
 doi:10.20944/preprints202101.0346.v1

The proposed method explains diabetic wounds with respect to wound tissue and ulcer location. Diabetic ulcers mostly occur under the foot and follow a similar pattern. A different diabetic wound occurs just below the ankle in Figure 8, which is misclassified as a lymphovascular wound. This kind of ulcer is hard to differentiate from lymphovascular wounds because of its location as lymphovascular wounds frequently occur at the ankle. The misclassification of a diabetic wound can also be the result of a large wound area that lymphovascular wounds tend to cover larger areas than diabetic ulcers.

Lymphovascular wounds are detected with high probability. There is a slightly lower probability of lymphovascular wound in Figure 9. The spread of the wound forms a line that looks like a surgical wound's scar. The darker part of the wound also looks like a diabetic ulcer. That's why the proposed model gives about 7 percent probability to each wound. Nonetheless, the proposed method highlights the important area for the lymphovascular wound correctly.

It is assumed that the pressure injury in Figure 10 is misclassified due to the size and the shape of the wound area. Pressure injury has a large wound area with surrounding damaged skin in general. As shown in Figure 10, the wound occurs under the foot, which is a common diabetic wound area, and also the wound area is smaller in comparison to the regular pressure injury wounds. That's why the proposed model misclassified the image of pressure injury wounds.

Figure 11 depicts a surgical wound, which is correctly classified with a probability of 63.4%. This surgical wound might be a result of a previous pressure injury which covers a large area. The vast spread of the wound causes this conclusion for the model. In addition to this, the model is confused with the edge of the white cloth, which causes a larger highlighted area. The darker and deeper wound in the middle might be the reason for the high diabetic wound percentage. On the other hand, surgical wounds tend to take a longer time to heal and might be converted to diabetic ulcers on diabetic patients. This is because diabetes causes wounds to heal slowly due to nerve damage and poor blood circulation. The classification performance of the model could be increased by collecting more data. This will strengthen the extraction of wounds' features in the training phase.



(a) Original Image **(b)** Importance map

Figure 8. *The probabilities of wound types. Diabetic: 29%, pressure injury: 14%, lymphovascular: 56%, surgical: 0% (Diabetic).*

 doi:10.20944/preprints202101.0346.v1

(a) Original Image **(b)** Importance map

Figure 9. *The probabilities of wound types: Lymphovascular: 87.8%, diabetic: 7.4%, pressure injury: 4.8%, surgical: 7.3% (Lymphovascular).*

(a) Original Image **(b)** Importance map

Figure 10. *The probabilities of wound types: Pressure injury: 27.3%, lymphovascular: 12%, surgical: 2.6%, diabetic: 58.2% (Pressure injury).*

(a) Original Image **(b)** Importance map

Figure 11. The probabilities of wound types: Surgical: 63.4%, diabetic: 19.4%, pressure injury: 15%, lymphovascular: 2.22% (Surgical).

4. Results and Discussion

The proposed model extracts features with convolutional networks from a pre-trained VGG16 network. The use of transfer learning accelerates training and produces efficient results, as seen in Figures 3-6. Performance metric evaluation of the model on diabetic wounds (with a precision of 0.72, recall of 0.71, and F1-score of 0.71) indicates that the model has limitations with feature identification for this wound type. This is especially evident with sparse datasets. Surgical wounds also have a low performance on the evaluation metrics where precision, recall, and F1-scores are 0.77, 0.70, 0.73, respectively. Precision recall and F1-scores of lymphovascular wounds are 0.87, 0.82, and 0.84, respectively. Pressure injury wound type has the highest precision, 0.96, a low recall score, 0.63, and an F1-score of 0.76. Lymphovascular and pressure injury wounds have good precision and F1-scores. The recall score of pressure injury wounds is low, which is an indicator that the proposed model has some difficulty in learning the features of pressure injury wounds.

The proposed model has the average of the precision at 0.83, the recall at 0.71, the F1-score at 0.76.

The second part of the model specialized in explaining why the model gives a specific output with a hybrid structure. This part extends the LIME technique using a heatmap model. The explainer of the CWC-XAI model is successful, while the classification part of the hybrid model has some limitations due to dataset size (a common problem in data-hungry deep learning models). The explainer provides visual cues through the use of a heatmap overlaid on wound images to indicate image regions identified by the AI model.

A clinician may eliminate certain wound types for consideration based on the location of the wound. For example, for a plantar foot ulcer, the doctor will likely eliminate sacral pressure injury wounds from the possible wound type list. This is why wound location is important, and an explanation of a wound type should indicate location information for a complete understanding. Diabetic wound type is explained via the corresponding deeper and darker damaged tissue size and location on toes. These features are stressed and shown in Figure 4. Lymphovascular wound features are highlighted and shown in Figure 5, where the size and texture of the damaged tissue are essential indicators. The explanation of the lymphovascular wound type is unexpected; its focus is on the border of the lesion and the adjacent areas instead of the whole lesion. This is another case that deep learning utilizes a non-intuitive search space that provides important information. Pressure injury wounds are explained via the wound tissue and the surrounding area of the wound, as seen in Figure 6. Pressure injury wounds often have a surrounding region of newly healed or damaged skin immediately adjacent to the larger wound. A surgical wound has more straightforward features to explain, such as postoperative scar and stitches.

Observations deduced from results of the CWC-XAI model are summarized below:

Observation 1: AI applications with XAI have a high potential in improving explainability and transparency in high-risk industries such as healthcare, where trust is key.

Observation 2: Quality and quantity of the training dataset and pre-processing affect model performance.

Observation 3: The limitation in the classification task is carried to the explanation part of the model.

Observation 4: The list of possible wound types is decreased significantly based on wound location.

Observation 5: Explainer has different approaches for each class, yet it uses a qualitative method to explain decisions.

Observation 6: Qualitative methods may explain AI models better to non-subject experts as model parameters and inputs alone are too numerous to be meaningful to non-data scientists.

Observation 7: Human reasoning tends to use both quantitative and qualitative methods, but if there is a hardship understanding quantitative methods, human reasoning could benefit highly from qualitative methods.

Observation 8: XAI has great potential to improve overall model performance by analyzing the effect and importance of features.

Observation 9: Non-expert users are often able to intuitively grasp the rationale behind class decisions made by the model.

5. Conclusions

This paper presents a use case of wound type classification in the healthcare domain using an explainable artificial intelligence model. The proposed model is used to augment decision-making through clinician guidance. Moreover, the proposed method reveals the underlying reason for a particular output by analyzing the relationship between input and output.

DNN using the transfer learning technique is utilized to predict the classification of four wound types, i.e., diabetic, lymphovascular, pressure injury, and surgical. The model

accepts an image as input and predicts the etiology of a chronic wound as output. It is discussed that trust is crucial for effective human interaction with machine learning systems and that explaining individual predictions is important in assessing trust. We used the XAI techniques identified here in a healthcare application to faithfully explain predictions of wound type classifications in an interpretable manner through the use of heatmaps. The proposed model extends the LIME technique with a heatmap method for better explainability. XAI techniques allow AI systems to cooperate with non-expert end-users. The AI and end-user give feedback to each other and conclude to a decision together by guiding a human, e.g., researcher or caregivers, during a classification task. It can also explain how a decision was made, tracing right back to the inner workings of the AI system. Transparency is crucial in developing caregiver confidence and improving wound treatment. In addition to transparency, new ways to implement a task will be discovered thanks to the distinctive approach of DL models.

This study demonstrated that explanations are useful for wound type classification in the healthcare domain, when assessing trust, developing new approaches to wound classification, and getting insights into predictions. The proposed hybrid model performs well on both chronic wound classification and explanation tasks. However, collecting more data will increase the classification performance further. Interpretation of the results obtained from the XAI module provides satisfactory information about the chosen wound type. Application of other XAI techniques such as Taylor Decomposition, Grad-CAM, and sensitivity analysis will enhance the overall trustworthiness of the model as well.

It is expected that this work can benefit researchers and caregivers, who work on the chronic wound management field in healthcare by providing an insight into the XAI potentials and availability in healthcare applications.

Author Contributions: Conceptualization, S.S., M.K., and O.G.; methodology, S.S., M.K., and O.G.; software, S.S., M.K., and O.G.; validation, S.S., M.K., E.W. , U.C., and O.G.; formal analysis, S.S., M.K., and O.G.; investigation, S.S., M.K., E.W., U.C., and O.G.; resources, S.S., M.K., and O.G.; data curation S.S., M.K., and O.G.; writing—original draft preparation, S.S. and M.K.; writing—review and editing, S.S., M.K., E.W., U.C., and O.G.; visualization, S.S. and O.G.; supervision, M.K., and O.G.; project administration, M.K., and O.G.. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Muggleton, S., 2014. Alan Turing and the development of Artificial Intelligence. AI Communications, 27(1), pp.3-10.
2. Villani, C., Bonnet, Y. and Rondepierre, B., 2018. For a meaningful artificial intelligence: Towards a French and European strategy. Conseil national du numérique.
3. Lu, H., Li, Y., Chen, M. et al. Brain Intelligence: Go beyond Artificial Intelligence. Mobile Netw Appl 23, 368–375 (2018).
4. OECD (2019), Artificial Intelligence in Society, OECD Publishing, Paris.
5. Ramesh, A.N., Kambhampati, C., Monson, J.R., and Drew, P.J., 2004. Artificial intelligence in medicine. Annals of the Royal College of Surgeons of England, 86(5), p.334.
6. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H. and Wang, Y., 2017. Artificial intelligence in healthcare: past, present, and future. Stroke and vascular neurology, 2(4), pp.230-243.
7. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)" in IEEE Access, vol. 6, pp. 52138-52160, 2018.
8. Castelvecchi, D., 2016. Can we open the black box of AI? Nature News, 538(7623), p.20.
9. Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B., 2017. What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923.
10. Gade K, Geyik SC, Kenthapadi K, Mithal V, Taly A. Explainable AI in industry. InProceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019 Jul 25 (pp. 3203-3204). ACM.
11. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, Chatila R. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020 Jun 1;58:82-115.

12. Došilović FK, Brčić M, Hlupić N. Explainable artificial intelligence: A survey. In2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO) 2018 May 21 (pp. 0210-0215). IEEE.
13. Ahmad MA, Eckert C, Teredesai A. Interpretable machine learning in healthcare. InProceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics 2018 Aug 15 (pp. 559-560).
14. Schmelzer, R., Understanding Explainable AI, Jul 23, 2019. Accessed on: Jun 1, 2020. [Online]. Available: https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/#5d112a887c9e
15. Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.
16. Mathews SM. Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review. Intelligent Computing-Proceedings of the Computing Conference 2019 Jul 16 (pp. 1269-1292). Springer, Cham.
17. Aghamohammadi M, Madan M, Hong JK, Watson I. Predicting Heart Attack through Explainable Artificial Intelligence. International Conference on Computational Science 2019 Jun 12 (pp. 633-645). Springer, Cham.
18. Monteath I, Sheh R. Assisted and Incremental Medical Diagnosis using Explainable Artificial Intelligence. XAI 2018.
19. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296. 2017 Aug 28.
20. Meske C, Bunde E. Using Explainable Artificial Intelligence to Increase Trust in Computer Vision. arXiv preprint arXiv:2002.01543. 2020 Feb 4.
21. S. J. Pan and Q. Yang, "A Survey on Transfer Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.
22. Dai, W., Yang, Q., Xue, G.R. and Yu, Y., 2007, June. Boosting for transfer learning. In Proceedings of the 24th international conference on Machine learning (pp. 193-200).
23. Torrey, L. and Shavlik, J., 2010. Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques (pp. 242-264). IGI Global.
24. Ge, L., Gao, J., Ngo, H., Li, K. and Zhang, A., 2014. On handling negative transfer and imbalanced distributions in multiple source transfer learning. Statistical Analysis and Data Mining: The ASA Data Science Journal, 7(4), pp.254-271.
25. Gunning, D., 2017. Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA).
26. Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B., 2017. What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923.
27. Doran, D., Schulz, S., and Besold, T.R., 2017. What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794.
28. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. Nature 577, 89–94 (2020)
29. Sheh, R. and Monteath, I., 2017. Introspectively assessing failures through explainable artificial intelligence. In IROS Workshop on Introspective Methods for Reliable Autonomy.
30. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I., 2019. Explainable ai for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610.
31. Magnani, L., 2019. AlphaGo, Locked Strategies, and Eco-Cognitive Openness. Philosophies, 4(1), p.8.
32. Hulstaert, L., Interpreting machine learning models, Feb 20, 2018. Accessed on: Jun 5, 2020. [Online]. Available: https://towardsdatascience.com/interpretability-in-machine-learning-70c30694a05f
33. Vilone, G. and Longo, L., 2020. Explainable Artificial Intelligence: a Systematic Review. arXiv preprint arXiv:2006.00093.
34. Ribeiro, M.T., Singh, S., and Guestrin, C., 2016. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.
35. Sumit, S., Local Interpretable Model-Agnostic Explanations (LIME) — the ELI5 way, Aug 14, 2019. Accessed on: Jun 15, 2020. [Online]. Available: https://medium.com/intel-student-ambassadors/local-interpretable-model-agnostic-explanations-lime-the-eli5-way-b4fd61363a5e
36. eKare, Inc., Accessed on: Jun 15, 2020. [Online]. Available: https://ekare.ai/.
37. Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
38. Struss, P., 1997. Model-based and qualitative reasoning: An introduction. Annals of Mathematics and Artificial Intelligence, 19(3-4), pp.355-381.
39. Fong, R.C., and Vedaldi, A., 2017. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3429-3437).