electronics

MDPI

Article

The Enlightening Role of Explainable Artificial Intelligence in Chronic Wound Classification

Salih Sarp [1], Murat Kuzlu [2], Emmanuel Wilson [3], Umit Cali [4,*] and Ozgur Guler [3]

[1] Department of Electrical and Computer Engineering, Virginia Commonwealth University, Richmond, VA 23284, USA; ssarp001@odu.edu
[2] Batten College of Engineering & Technology, Old Dominion University, Norfolk, VA 23529, USA; mkuzlu@odu.edu
[3] eKare Inc., Fairfax, VA 22031, USA; ewilson@ekareinc.com (E.W.); oguler@ekareinc.com (O.G.)
[4] Department of Electric Power Engineering, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
* Correspondence: umit.cali@ntnu.no

Abstract: Artificial Intelligence (AI) has been among the most emerging research and industrial application fields, especially in the healthcare domain, but operated as a black-box model with a limited understanding of its inner working over the past decades. AI algorithms are, in large part, built on weights calculated as a result of large matrix multiplications. It is typically hard to interpret and debug the computationally intensive processes. Explainable Artificial Intelligence (XAI) aims to solve black-box and hard-to-debug approaches 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 to extract additional knowledge that can also be interpreted by non-data-science experts, such as medical scientists and physicians. 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

Citation: Sarp, S.; Kuzlu, M.; Wilson, E.; Cali, U.; Guler, O. The Enlightening Role of Explainable Artificial Intelligence in Chronic Wound Classification. *Electronics* **2021**, *10*, 1406. https://doi.org/10.3390/electronics10121406

Academic Editors: Hüseyin Kusetogullari, Turgay Celik, Chafik Samir, Amir Yavariabdi, Antonio Orlandi and Byung-Gyu Kim

Received: 10 April 2021
Accepted: 7 June 2021
Published: 11 June 2021

1. Introduction

After alternating between periods of great passion and setback [1], AI has found its place as a critical component of growth in a variety of applications [2]. These applications range from diagnostic decision assistants in healthcare, safety-critical systems in autonomous vehicles, and long-term financial investment planning, and benefit from these breakthroughs [3].

AI is capable of analyzing complex data and exploiting non-intuitive approaches to derive meaningful relationships [4]. Healthcare applications based on AI are utilized in early detection, diagnosis, treatment, as well as outcome prediction and prognosis evaluation [5]. The barrier that stands in the way of AI applications is sourced from the lack of transparency and black-box nature that cannot be explained directly [6]. The black-box nature of AI systems could be explained as follows. 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 [7]. This is why an explainable and understandable glass-box approach should be taken to enable transparent, trustable, and retraceable AI applications [8]. Chronic wound management, which is one of the important fields in healthcare, also requires explainable AI models. In this study, AI techniques are applied to the classification of chronic wounds, i.e., diabetic ulcers, lymphovascular, surgical, and pressure injury.

The Explainable Artificial Intelligence (XAI) term is coined to provide transparency and guided inference in understanding the decision-making processes of the AI system [9]. The study in [10] provides a comprehensive review of XAI in terms of concepts, taxonomies, opportunities, and challenges, as well as a discussion on adopting XAI techniques to image processing. The study in [11] summarizes the recent developments in XAI and its connection with artificial general intelligence, as well as identified trust-related problems of AI applications. The study in [12] examines the state of AI-based FDA-approved medical devices and algorithms. Although millions of dollars funded medical AI research in 2019, only ten (10) medical devices have been approved by the FDA. The authors in [13] present a comparative analysis of approved AI and ML medical devices. The approved devices are being used mainly in radiology, and a few are qualified as high-risk devices. The acceptance of AI is still low amongst medical practitioners with various matters related to trustworthiness and reliability [14]. Authors in [15] identified nuances, challenges, and requirements for the design of interpretable and explainable machine learning models and systems in healthcare and described how to choose the right interpretable machine learning algorithm. Conventional black-box AI systems are turned into glass-box systems with the help of XAI techniques which provide data about the intermediate steps of the inference process [16,17]. 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 [18] 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 [19] 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 [20], 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 satisfactory rates. Authors in [21] developed an assisted and incremental medical diagnosis system using XAI, which allows the interaction between the physician (i.e., human agent) and the AI agent. Authors in [22] investigated the problem of explainability in AI in the medical domain where wrong system decisions 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 [23] 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 the classification of chronic wounds, i.e., diabetic ulcer, lymphovascular, surgical, and pressure injury. Objectives of the study are:

- Build a wound type classification model using deep learning and transfer learning methods.
- Showcase an approach to make common AI models more transparent and explainable to understand the results and gain trust into the AI model.
- Utilize readily available AI neural networks to show that more transparency or explainability can be introduced to a variety of commonly available models, such as transfer learning.
- Apply XAI methods to convert complex black-box AI systems to more understandable glass box AI systems that aim to provide a look into the internal decision-making mechanics to give the user the ability to follow the reasoning behind the AI models' prediction.
- Provide insights into the complex decision-making processes of an AI system in the field of healthcare applications, especially chronic wound type classification.

2. Methodology

This section discusses the methodology of transfer learning for the wound type classification and XAI for providing transparency to the classification task as well as the overall model pipeline.

2.1. Transfer Learning

Predictions on new data utilizing data distributions and statistical properties of a previously trained model are called transfer learning [24]. The same distribution of the training and the testing dataset is needed for traditional machine learning models [25]. 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.

The transfer learning application comprises two steps, (i) feature extraction, and (ii) fine-tuning. The pre-trained network will extract meaningful features from new data samples, with a final classifier added on top of the pre-trained network to do classification tasks in the target domain. The pre-trained network masters feature extraction task with convolutional networks. The second step is fine-tuning through freezing and unfreezing some of the top layers from the pre-trained model to train for higher performance jointly. ResNet [26], EfficientNet [27], and VGG16 (Very Deep Convolutional Neural Networks for Large-Scale Image Recognition) [28] networks are a few of the successful DL models for classification tasks. In this study, transfer learning is utilized with VGG16 architecture in order to utilize its object detection capabilities. Its architecture is 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 [29] database. The network is initialized with random weights before the training [30]. The pre-trained convolution layers of the VGG16 architecture are kept frozen, and only fully connected output layers after convolutional 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 is kept unfrozen. The last convolution layer and fully connected layers are trained together to fine-tune the model, i.e., deep neural networks (DNN). The weights of convolution layers from the VGG16 are transferred to utilize their feature extraction skills. The training of the last convolution layer provides the fine-tuning necessary to obtain better classification results.

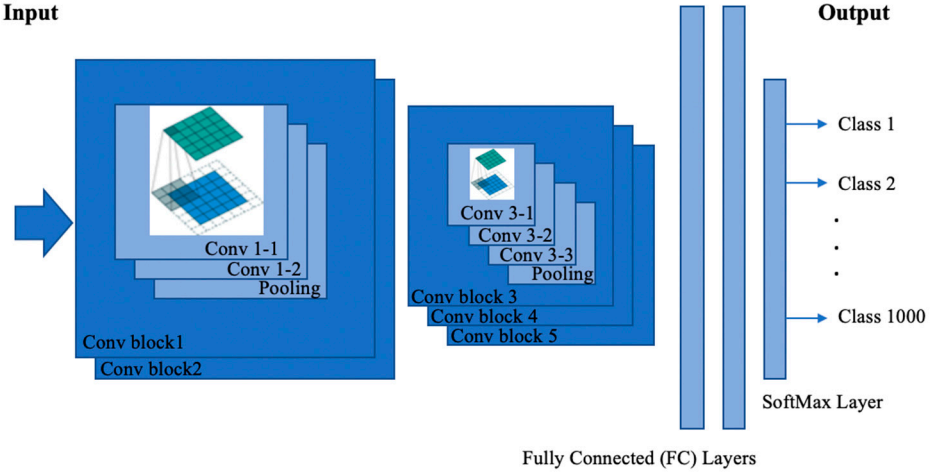


Figure 1. VGG16 architecture.

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 [31]. 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 [32]. There are

two kinds of approaches to explain an AI model; (i) the comprehensible, and (ii) the interpretable model. Comprehensible models are explained with post-hoc explainability approaches. Classical machine learning methods (e.g., regression models and decision trees) are interpretable models as these reveal greater transparency when compared to convolutional networks [33]. Inner workings of machine learning models might be complicated and hard to interpret, yet their efficiency and accuracy are higher than human performance in many cases [34]. This improved efficiency and accuracy are the main reasons why we need to comprehend the inner workings of machine learning models.

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 [35].

Tree-based models have individually meaningful features, with tabular-style datasets used in these models. The connection of tree-based models to the training data results in greater interpretability with local explanations in comparison to linear regression models [36].

DL is a relatively new research field compared to classical machine learning models. 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 the image and text-based models, respectively. The gradient-based method brings attention to important regions in the input image in the backward pass. The attention-based method trains attention weights, which determine how much each of the elements is in the final output [37].

Generalized explainable AI methods are designed to treat any machine learning model as a black-box with inputs and some outputs [38]. One of these methods is Local Interpretable Model-Agnostic Explanations (LIME) [39]. LIME finds the statistical connection between input and model prediction by training local surrogate models on perturbed inputs instead of training them globally [40]. It provides both an explanation of an instance by an interpretable representation as well as visualization. This study provides the explainability and transparency of chronic wound classification using transfer learning implementation with Keras and XAI methods.

2.3. Model Pipeline

The proposed model architecture consists of two main parts, i.e., classification, and explanation. In the first part of the process, the chronic wound images are classified into four categories, i.e., diabetic, lymphovascular, pressure injury, and surgical. This part of the model employs a pre-trained VGG16 network, i.e., transfer learning, which is capable of extracting features using 13 convolution (conv) layers. These layers are already loaded with pre-trained weights using the ImageNet dataset that is publicly available. The last three fully connected (FC) layers and the softmax layer is trained with the chronic wound dataset from the ground up to provide weights for the classification of chronic wounds. After training the classification part of the model with these steps, images are fed to the explainable AI part of the model, where the LIME XAI tool and heatmap are utilized for the explanation. The process of classification and explanation of chronic wound images is illustrated in Figure 2. The input wound image is simply classified by the model consisting of transfer learning and DNN and then explained with an XAI tool, i.e., LIME, and heatmap for providing transparency to the classification.

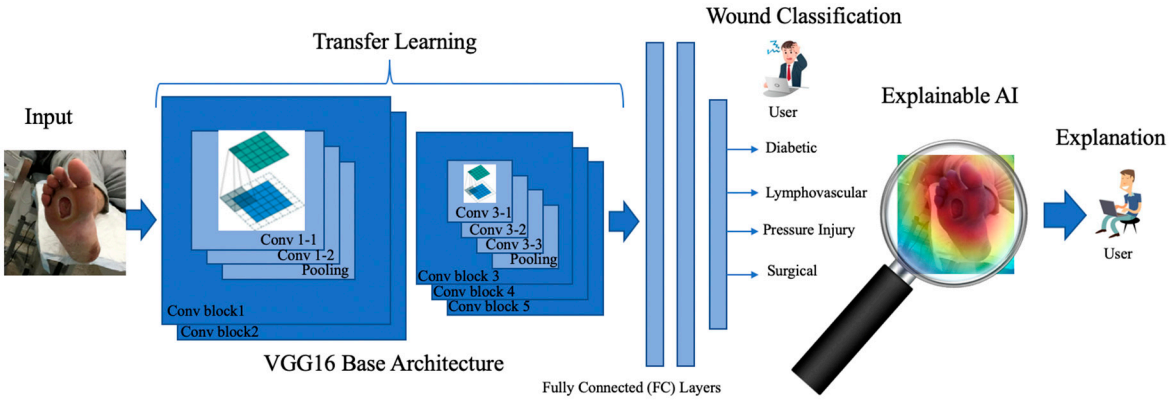


Figure 2. Wound classification model with transfer learning, DNN, and explainable AI tool.

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

This section discusses data collection, data pre-processing, and the test environment. 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, respectively.

3.1. Data Collection

The chronic wound data repository, which includes diabetic, lymphovascular, pressure injury, and surgical wound types, are collected from the eKare Inc. data repository and was anonymized for patient privacy [41]. eKare Inc. specializes in wound management, with its services used by many hospitals and wound clinics for patient/wound management. A total of 8690 wound images were chosen by an MD specialized in wound care to represent the aforementioned wound types. The dataset comprises 1811 diabetic, 2934 lymphovascular, 2299 pressure injuries, and 1646 surgical wound images.

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 the correct classification of underlying chronic wound etiology. This validated classification serves as the clinical ground-truth. Wound images are then hand-labeled for 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 such as mirroring, rotation, and horizontal flip are used to increase dataset size and maintain class balance. The dataset, 8690 images in total, was split into training and test sets comprising 6520 and 2170 images, respectively. The collected data was pre-processed to increase data quality. This includes formatting, rescaling, and normalization of the images. Images were scaled to 224 $\times$ 224 pixels and normalized for a faster learning process.

3.3. Environment

The proposed model was implemented using the 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, 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 h. We used a constant learning rate of 0.001 for the "RMSprop" optimizer for the training.

3.4. Validation

Validation was done using the confusion matrix shown in Table 1. Precision gives the ratio of correctly classified wound types over total positive wound type predictions. Recall is a measure of how many of the positive wounds are correctly classified. This metric checks predictions in the eye of true labels. A high recall value relates to the identification of more true positive, and therefore, fewer incorrectly classified samples. Interestingly, both of these metrics could be high, yet the model could still underperforms. This is why a third metric is utilized to characterize the model performance. F1-score is a hybrid measurement that brings together both precision and recalls for a better evaluation.

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 Equations (1)–(3) below.

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

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

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

(ROC) curve and area under the curve (AUC) are also used as performance measures and shown in Figure 3. Higher AUC values indicate the classification capability of the proposed model. The X-axis of the ROC curve is recall, and Y-axis is the false positive rate (FPR) which is given in Equation (4) below.

$$\text{FPR} = \frac{\text{False Positive}}{\text{True Negative} + \text{False Positive}} \quad (4)$$

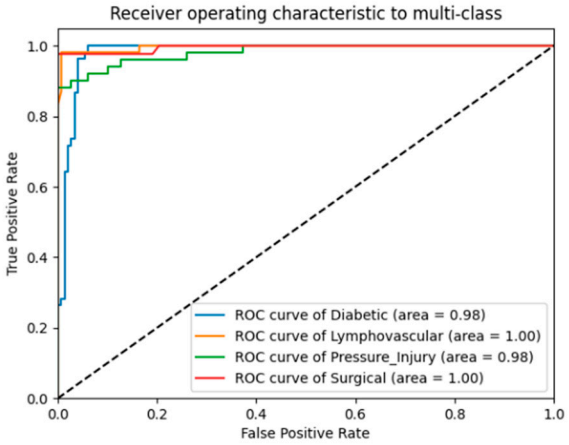


Figure 3. ROC curve of the wound classification.

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 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 to 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, re-call, and precision (features extracted from the confusion matrix). Precision, recall, and F1-scores of each wound type, and their averages, are compared in Table 2.

Table 2. Classification performance evaluation of the proposed model.

Model	Precision	Recall	F1-Score
Diabetic	0.85	1.00	0.92
Lymphovascular	0.95	0.98	0.96
Pressure Injury	1.00	0.86	0.92
Surgical	1.00	0.91	0.95
Average	0.95	0.94	0.94

Higher precision values of lymphovascular, surgical, and pressure injury wound types indicate the model performed very well with these wound types, whereas pressure injuries were harder to diagnose (low recall score for pressure injury wounds). This means that some pressure injury wounds are not learned or are similar to another wound type and misclassified by the model. Lymphovascular wounds have one of the highest recall scores among all wound types, which reveals that the 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 F1 scores, but have low recall scores. Hence our model is likely to classify a surgical wound as diabetic. The recall of diabetic wound types is pretty high, and it has one of the lowest F1 scores, which is a result of low precision. The ROC curve and AUC results are depicted in Figure 3. Lymphovascular and surgical wounds have the highest AUC values, whereas diabetic and pressure injury suffers from low precision (diabetic) and recall (pressure injury).

As AI-based products provide efficiency and automation, AI becomes very popular in low-risk fields, such as agriculture, customer services, and manufacturing. However, applications of AI remain limited in high-risk domains such as health care, as trust is critical in medical practice [14]. Reliability issues concerning patients and medical practitioners, as well as regulations, hinder the adoption of AI-based systems [12]. Understanding the rationale behind model predictions would certainly help users decide when to trust or not to trust their predictions.

A deep neural network using the transfer learning technique was trained using chronic wound images to predict the 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 aspect [42]. The rich model feature-set is too numerous and difficult to interpret directly, yet by facilitating a guided qualitative approach, human reasoning can be augmented with additional model data [43]. Another significant property that a reliable explainer should have is local faithfulness. Local faithfulness is achieved by characterizing the response of a local function with a range of adjacent inputs [44].

In this study, the DNN model with transfer learning and extended XAI technique is used to provide explainability and transparency for wound image classifiers by visually indicating what particular class is estimated for various model regions. The proposed

model forms a hybrid XAI framework through the use of LIME and heatmap proposals. LIME architecture using superpixels is implemented similar to the study in [42]. LIME provides a set of correlated and connected pixels which are used as input to the heatmap method. The proposed model provides a focus on the classification task through the use of a heatmap. Medical practitioners often conceptualize the clinical problem based on their knowledge acquired in medical school, as well as clinical experience. The heatmap approach is a fairly naïve method of raising focus to different image regions based on the model. The basic intuition with the use of the heatmap is that by drawing focus to certain image regions, practitioners will narrow their attention to regions where the heatmap data correlates with their medical intuition. 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%, shown in Figure 4. Figure 4b highlights the model's focused area for classification tasks in the wound image with an importance map as an explanation.

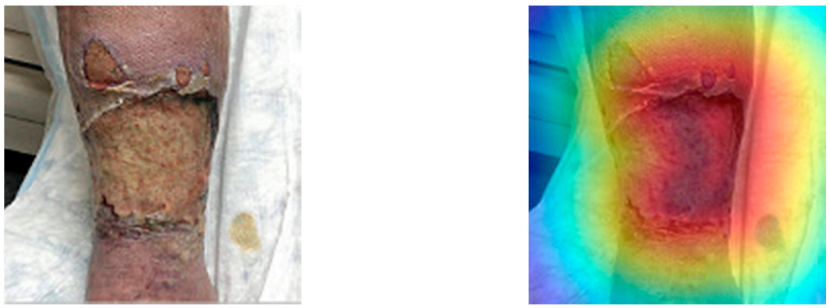
(**a**) Original wound image (**b**) Importance map

Figure 4. Original lymphovascular wound image and its explanation using heatmap.

Figures 5–8 show images of diabetic, lymphovascular, pressure injury, and surgical wounds. Each wound type has a respective heatmap highlighting 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: 0.56%) and lymphovascular wound is predicted at 100% (Diabetic: 0%, pressure injury: 0%, surgical: 0%) in Figures 5 and 6, respectively. The low diabetic wound classification probability can be increased with additional data to amplify feature extraction of diabetic wounds during training.

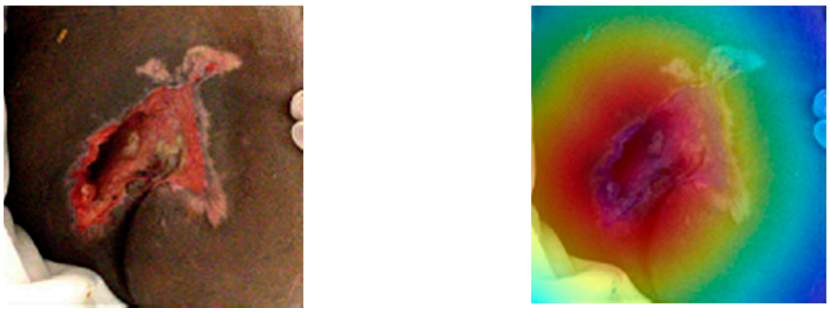
(**a**) Original Image (**b**) Importance map

Figure 5. The probabilities of wound types: Diabetic: 95.36%, pressure injury: 4.07%, lymphovascular: 0.01%, surgical: 0.56% (Diabetic).



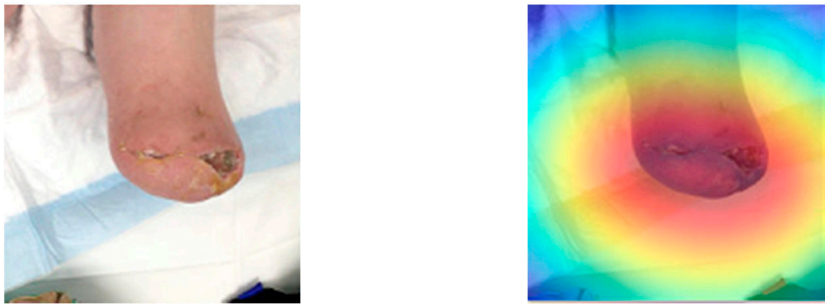
(**a**) Original Image (**b**) Importance map

Figure 6. The probabilities of wound types: Lymphovascular: 100%, diabetic: 0%, pressure Injury: 0%, surgical: 0% (Lymphovascular).



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

Figure 7. The probabilities of wound types: Pressure Injury: 100%, lymphovascular: 0%, surgical: 0%, diabetic: 0% (Pressure Injury).



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

Figure 8. The probabilities of wound types: Surgical: 99.91%, diabetic: 0.05%, pressure injury: 0.03%, lymphovascular: 0.01% (Surgical).

Probabilities of wound classification are very high for Figure 7, i.e., pressure injury wound at 100% (Lymphovascular: 0%, surgical: 0%, diabetic: 0%), and for Figure 8, i.e., surgical wound at 99.91% (Diabetic: 0.05%, pressure injury: 0.03%, lymphovascular: 0.01%).

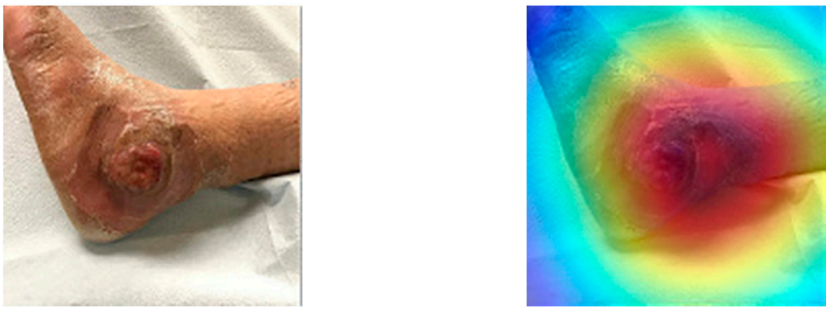
Figure 5a,b show explanations of the most important features that contribute to the prediction. Like Figure 5a,b, Figure 6a,b shows explanations and map features with the highest contribution to prediction for lymphovascular classification. Both figures provide insights as to why the wound type was predicted to be diabetic or lymphovascular. Focus on the diabetic wound includes surrounding wound tissues and toes, with the shape of the ulcer and its proximity to toes as the explanations of the diabetic foot ulcer.

The lymphovascular wound, as seen in Figure 5a, 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 pressure injury wound explainer focuses on the wounded area and indicates the correct placement of the wound, shown in Figure 7b. In Figure 8, 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.

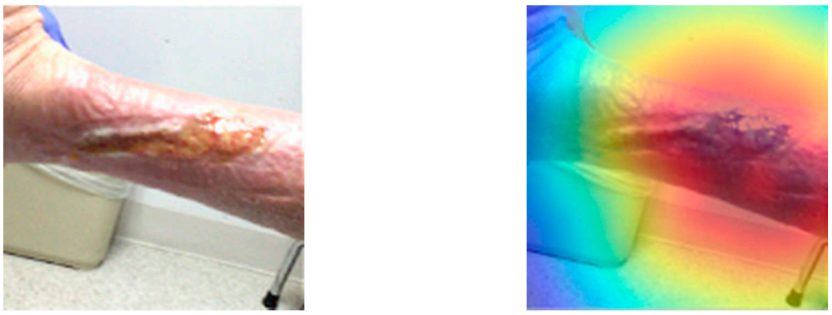
The proposed method explains diabetic wounds with respect to wound tissue and ulcer location. Diabetic ulcers mostly occur under the foot and follow a similar pattern. A different diabetic wound occurs just below the ankle in Figure 9, which is misclassified as a lymphovascular wound. This kind of ulcer is hard to differentiate from lymphovascular wounds because of its location, as lymphovascular wounds frequently occur at the ankle. Misclassification of a diabetic wound can also be the result of a large wound area, wherein lymphovascular wounds typically cover larger areas than diabetic ulcers.



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

Figure 9. The probabilities of wound types. Diabetic: 29%, pressure injury: 14%, lymphovascular: 56%, surgical: 1% (Diabetic).

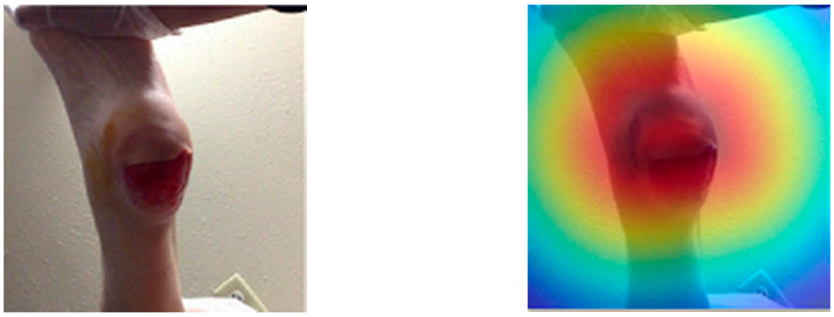
Lymphovascular wounds are detected with high probability. There is a slightly lower probability of a lymphovascular wound in Figure 10. The spread of the wound forms a line that looks like a surgical wound's scar. The darker part of the wound also looks like a diabetic ulcer. That's why the proposed model gives about 7 percent probability to each wound. Nonetheless, the proposed method highlights the important area for the lymphovascular wound correctly.



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

Figure 10. The probabilities of wound types: Lymphovascular: 80.8%, diabetic: 7.4%, pressure injury: 4.5%, surgical: 7.3% (Lymphovascular).

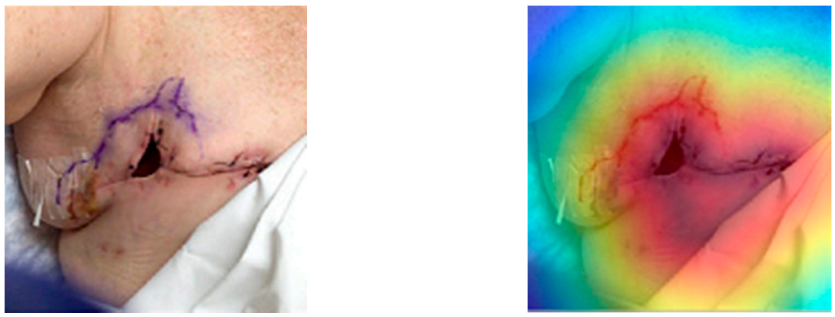
It is assumed that the pressure injury wound in Figure 11 is misclassified due to the size and the shape of the wound area. Pressure injury typically has a large wound area with surrounding damaged skin. As shown in Figure 11, the wound occurs under the foot, which is a common diabetic wound area, and also the wound area is smaller in comparison to the regular pressure injury wounds. These comprise the reasons why the proposed model misclassified the image of pressure injury wounds.



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

Figure 11. The probabilities of wound types: Pressure injury: 27.3%, lymphovascular: 12%, surgical: 2.5%, diabetic: 58.2% (Pressure injury).

Figure 12 depicts a surgical wound, which is correctly classified with a probability of 63.4%. This surgical wound might be the result of a previous pressure injury that covered a larger 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 may convert to diabetic ulcers in diabetic patients. Model classification performance could be increased by collecting more data as this will strengthen the extraction of wound features in the training phase.



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

Figure 12. The probabilities of wound types: Surgical: 63.4%, diabetic: 19.4%, pressure injury: 15%, lymphovascular: 2.2% (Surgical).

5. 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 shown in Figures 4–8. Performance metric evaluation of the model on diabetic wounds (with a precision of 0.85, recall of 1.00, and F1-score of 0.92) indicate that the model has limitations with feature identification for this wound type. This is especially evident with sparse datasets. Surgical wounds have a fair performance on the evaluation metrics where precision, recall, and F1-scores are 1.00, 0.91, 0.95, respectively. Precision, recall, and F1 scores of lymphovascular wounds are 0.95, 0.98, and 0.96, respectively. Pressure injury wound type has one of the highest precisions, 1.00, a low recall score, 0.86, and an F1-score of 0.92. Surgical and pressure injury wounds have good precision and low recall 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 precision at 0.95, the recall at 0.94, the F1-score at 0.94. The ROC curve and the AUC provide a visualization related to the performance of the model on the classification task. The performance of the model could be improved with a larger training dataset [45] and fine-tuning the hyperparameters [46].

The second part of the model is specialized in explaining why the model gives a specific output with a hybrid structure. This part extends the LIME technique using a heatmap model. Heatmap is used as a tool to draw focus to image regions based on work done with the intuition being that practitioners will take less time under guidance. The explainer of the proposed model is successful, while the classification part of the hybrid model could be further improved with additional data (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, in the case of a plantar foot ulcer, a 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 also 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 5. Lymphovascular wound features are highlighted and shown in Figure 6, where the size and texture of the damaged tissue are essential indicators. 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 whereby deep learning utilizes a non-intuitive search space that provides important information. Pressure injury wounds are explained via wound tissue and the surrounding wound area, as seen in Figure 7. 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 the results of the proposed model are summarized below:

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

Observation 2: Limitation in the classification task is carried to the explanation part of the model.

Observation 3: The list of possible wound types is decreased significantly based on wound location.

Observation 4: Explainer has different approaches for each class, yet it uses a qualitative method to explain decisions.

Observation 5: 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-experts.

Observation 6: Given hardships in understanding quantitative methods, human reasoning can be augmented through qualitative methods.

Observation 7: XAI has great potential to improve overall model performance by analyzing the effect and importance of features.

Observation 8: Non-expert users are often able to intuitively grasp the rationale behind class decisions made by the model.

Observation 9: AI decision-making processes might be unanticipated, yet they can provide insights and improve how we handle certain tasks through a bottom-up approach.

6. 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. This study intends to showcase an approach to make common AI models more transparent and explainable to understand the results and gain trust into the AI model. By utilizing readily available AI neural networks, it can be shown that more transparency or explainability can be introduced to a variety of commonly available models, such as transfer learning.

DNN using the transfer learning technique is utilized to predict the classification of four wound types: 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 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 each other feedback to arrive at 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 back to the inner workings of the AI system. Transparency is crucial in developing caregiver confidence and improving wound treatment.

This study demonstrated that explanations are useful for wound type classification in the healthcare domain, when assessing trust, to develop new approaches to wound classification and prediction insights. The proposed hybrid model performs well on both chronic wound classification and explanation tasks. Collecting additional data will increase 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 in the chronic wound management field in healthcare by providing insights 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.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank the anonymous reviewers for their contribution to this paper.

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

References

1. Villani, C.; Bonnet, Y.; Rondepierre, B. *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*; Conseil National du Numérique, 2018. Available online: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf (accessed on 15 June 2020).
2. Lu, H.; Li, Y.; Chen, M.; Kim, H.; Serikawa, S. Brain Intelligence: Go beyond Artificial Intelligence. *Mob. Netw. Appl.* **2018**, *23*, 368–375. [CrossRef]
3. OECD. *Artificial Intelligence in Society*; OECD Publishing: Paris, France, 2019.
4. Ramesh, A.N.; Kambhampati, C.; Monson, J.R.; Drew, P.J. Artificial intelligence in medicine. *Ann. R. Coll. Surg. Engl.* **2004**, *86*, 334. [CrossRef]
5. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present, and future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. [CrossRef]
6. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
7. Castelvecchi, D. Can we open the black box of AI? *Nat. News* **2016**, *538*, 20. [CrossRef]
8. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What do we need to build explainable AI systems for the medical domain? *arXiv* **2017**, arXiv:1712.09923.

9. Gade, K.; Geyik, S.C.; Kenthapadi, K.; Mithal, V.; Taly, A. Explainable AI in industry. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 25 July 2019; ACM: New York, NY, USA; pp. 3203–3204.
10. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
11. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable artificial intelligence: A survey. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21 May 2018; IEEE: Piscataway, NJ, USA; pp. 0210–0215.
12. Benjamens, S.; Dhunnoo, P.; Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *NPJ Digit. Med.* **2020**, *3*, 1–8. [CrossRef] [PubMed]
13. Muehlematter, U.J.; Daniore, P.; Vokinger, K.N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. *Lancet Digit. Health* **2021**. [CrossRef]
14. Park, C.W.; Seo, S.W.; Kang, N.; Ko, B.; Choi, B.W.; Park, C.M.; Chang, D.K.; Kim, H.; Kim, H.; Lee, H.; et al. Artificial Intelligence in Health Care: Current Applications and Issues. *J. Korean Med Sci.* **2020**, *35*, 379. [CrossRef]
15. Ahmad, M.A.; Eckert, C.; Teredesai, A. Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Washington, DC, USA, 15 August 2018; pp. 559–560.
16. Rai, A. Explainable AI: From black box to glass box. *J. Acad. Mark. Sci.* **2020**, *48*, 137–141. [CrossRef]
17. Schmelzer, R. Understanding Explainable AI. 2019. Available online: https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/#5d112a887c9e (accessed on 1 June 2020).
18. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22 October 2017.
19. Mathews, S.M. Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review. In Proceedings of the Intelligent Computing-Proceedings of the Computing Conference, London, UK, 16–17 July 2019; Springer: Cham, Switzerland, 2019; pp. 1269–1292.
20. Aghamohammadi, M.; Madan, M.; Hong, J.K.; Watson, I. Predicting Heart Attack through Explainable Artificial Intelligence. In Proceedings of the International Conference on Computational Science, Faro, Portugal, 12–14 June 2019; Springer: Cham, Switzerland, 2019; pp. 633–645.
21. Monteath, I.; Sheh, R. Assisted and incremental medical diagnosis using explainable artificial intelligence. In Proceedings of the 2nd Workshop on Explainable Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 104–108.
22. Samek, W.; Wiegand, T.; Müller, K.R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv* **2017**, arXiv:1708.08296.
23. Meske, C.; Bunde, E. Using Explainable Artificial Intelligence to Increase Trust in Computer Vision. *arXiv* **2020**, arXiv:2002.01543.
24. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
25. Dai, W.; Yang, Q.; Xue, G.R.; Yu, Y. Boosting for transfer learning. In Proceedings of the 24th International Conference on Machine Learning, Corvalis, OR, USA, 20–24 June 2007; pp. 193–200.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; PMLR. pp. 6105–6114.
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
29. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
30. He, K.; Wang, Y.; Hopcroft, J. A Powerful Generative Model Using Random Weights for the Deep Image Representation. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 631–639.
31. Sarp, S.; Kuzlu, M.; Cali, U.; Elma, O.; Guler, O. An Interpretable Solar Photovoltaic Power Generation Forecasting Approach Using an Explainable Artificial Intelligence Tool. In Proceedings of the 2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington DC, USA, 16–18 February 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–5.
32. Goebel, R.; Chander, A.; Holzinger, K.; Lecue, F.; Akata, Z.; Stumpf, S.; Kieseberg, P.; Holzinger, A. Explainable ai: The new 42? In Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Hamburg, Germany, 27–30 August 2018; Springer: Cham, Switzerland, 2018; pp. 295–303.
33. Doran, D.; Schulz, S.; Besold, T.R. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv* **2017**, arXiv:1710.00794.
34. McKinney, S.M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.S.; Darzi, A.; et al. International evaluation of an AI system for breast cancer screening. *Nature* **2020**, *577*, 89–94. [CrossRef] [PubMed]

35. De, T.; Giri, P.; Mevawala, A.; Nemani, R.; Deo, A. Explainable AI: A Hybrid Approach to Generate Human-Interpretable Explanation for Deep Learning Prediction. *Procedia Comput. Sci.* **2020**, *168*, 40–48. [CrossRef]
36. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. Explainable ai for trees: From local explanations to global understanding. *arXiv* **2019**, arXiv:1905.04610. [CrossRef]
37. Hulstaert, L. Interpreting Machine Learning Models. 2018. Available online: https://towardsdatascience.com/interpretability-in-machine-learning-70c30694a05f (accessed on 5 June 2020).
38. Vilone, G.; Longo, L. Explainable Artificial Intelligence: A Systematic Review. *arXiv* **2020**, arXiv:2006.00093.
39. Ribeiro, M.T.; Singh, S.; Guestrin, C. Model-agnostic interpretability of machine learning. *arXiv* **2016**, arXiv:1606.05386.
40. Sumit, S. Local Interpretable Model-Agnostic Explanations (LIME)—The ELI5 Way. 2019. Available online: https://medium.com/intel-student-ambassadors/local-interpretable-model-agnostic-explanations-lime-the-eli5-way-b4fd61363a5e (accessed on 15 June 2020).
41. eKare, Inc. Available online: https://ekare.ai/ (accessed on 15 June 2020).
42. Ribeiro, M.T.; Singh, S.; Guestrin, C. 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, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA; pp. 1135–1144.
43. Struss, P. Model-based and qualitative reasoning: An introduction. *Ann. Math. Artif. Intell.* **1997**, *19*, 355–381. [CrossRef]
44. Fong, R.C.; Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; IEEE: Piscataway, NJ, USA; pp. 3429–3437.
45. Sarp, S.; Kuzlu, M.; Pipattanasomporn, M.; Guler, O. Simultaneous wound border segmentation and tissue classification using a conditional generative adversarial network. *J. Eng.* **2021**. [CrossRef]
46. Sarp, S.; Kuzlu, M.; Wilson, E.; Guler, O. WG2AN: Synthetic wound image generation using generative adversarial network. *J. Eng.* **2021**. [CrossRef]