dHealth 2019 – From eHealth to dHealth
D. Hayn et al. (Eds.)

doi:10.3233/978-1-61499-971-3-192

Photographic LVAD Driveline Wound Infection Recognition Using Deep Learning

Noël LÜNEBURG[a,1], Nils REISS[b], Christina FELDMANN[c], Pim van der MEULEN[a], Michiel van de STEEG[a], Thomas SCHMIDT[b], Regina WENDL[c], Sybren JANSEN[a]
[a]*Target Holding, Groningen, The Netherlands*
[b]*Schüchtermann-Schiller'sche Kliniken, Bad Rothenfelde, Germany*
[c]*Hannover Medical School, Department for Cardiothoracic, Transplantation and Vascular Surgery, Hannover, Germany*

Abstract. The steady increase in the number of patients equipped with mechanical heart support implants, such as left ventricular assist devices (LVAD), along with virtually ubiquitous 24/7 internet connectivity coverage is motive to investigate and develop remote patient monitoring. In this study we explore machine learning approaches to infection severity recognition on driveline exit site images. We apply a U-net convolutional neural network (CNN) for driveline tube segmentation, resulting in a Dice score coefficient of 0.95. A classification CNN is trained to predict the membership of one out of three infection classes in photographs. The resulting accuracy of 67% in total is close to the measured expert level performance, which indicates that also for human experts there may not be enough information present in the photographs for accurate assessment. We suggest the inclusion of thermographic image data in order to better resolve mild and severe infections.

Keywords. heart assist device, driveline infection, infection classification, convolutional neural network

1. Introduction

An increasing number of patients with heart failure classified as severe according to the New York Heart Association (NYHA) Classification [1], are treated with a mechanical support implant. This may either be for the period while waiting for the heart transplantation, or as a permanent solution, the so-called destination therapy [7]. A left ventricular assist device (LVAD) is a pumping device implanted onto the heart, taking over the main pump function of the left ventricle while the heart is still functional at a low percentage. The device relies on a permanent electrical connection to a control module and battery pack situated on the outside of a patient's body, through a driveline tube. The control module collects device operation data which is a valuable opportunity for data exchange and thus early detection of problems.

The driveline exit site is a delicate location, requiring continuous wound treatment and wound dressing, the latter to be renewed typically once in five days. Driveline

[1] Corresponding Author: Noël Lüneburg, Target Holding, Atoomweg 6B Groningen, The Netherlands, E-Mail: noel.luneburg@target-holding.nl.

infections occur frequently because the driveline exit site creates a conduit for the entry and proliferation of bacteria. This is one of the most severe adverse events for the patient, leading to the necessity of surgical wound revision or even the replacement of the assist device implant [8]. Driveline infection is defined as an infection affecting the soft tissues around the driveline outlet, accompanied by redness, warmth, and purulent discharge.

Telemonitoring of driveline exit sites can provide early detection of these symptoms and can aid in the remote diagnosis of relevant driveline infections. The majority of LVAD patients have positive reactions towards telemonitoring [9]. Photographs of the driveline exit site, taken by caregivers or patients themselves with their mobile devices during renewal of the wound dressing, are sent through a mobile application to the physician in charge in the patient's clinic. The image will be reviewed in combination with any available device data, clinical data and the accompanying patient-update on their well-being or quality of life. The aim is to prevent patients from having to travel to their clinics for check-ups too often, or to consult their local general practitioner, but even more so not to miss out on the early detection of an upcoming adverse event. Right now the state of the art is that patients are seen by their clinics once every three months, without any visual monitoring in between.

Before deep learning was widely accepted as a machine learning method in the image processing field, the support vector machine (SVM) and multi-layered perceptron (MLP) were popular choices for computer aided image analysis in the domain of photographic imaging [11] as well as non-photographic medical imaging [12]. Deep learning has been applied to skin cancer classification supported by a large data set [10] in which a deep CNN matched (and even outperformed in certain configurations) dermatologists in classification accuracy.

In the following sections we describe three applications of deep learning which support the diagnosis procedure by automatically predicting the presence and severity of driveline infections based on patient photographic data. This can be executed 'on the fly' and will not add significant transit time to the images. The physician-in-charge then receives the images with a severity indication, and in particular a warning sign in case of a recognized severe infection.

2. Methods

The data set we worked on for this study consists of 745 general photographs from a total of 61 patients, taken and provided in pseudonymized format by Schüchtermann-Schiller'sche Kliniken and Hannover Medical School. The photographs had been taken and stored for documentation, without being further processed for some time. Photographs were taken from various positions and lack consistency in lighting. In addition, photographs can be out of focus or show signs of camera motion, and part of the wound area can be obstructed by dressing. These conditions might apply to future images taken by patients as well and we are prepared to handle this automatically.

732 out of the 745 photographs are labelled as belonging to one of the following three classes: *no infection, mild infection, severe infection*. In regular operations, labels are assigned by clinical experts based on features such as presence of bacteria, odour and warmth, in addition to visual features on the surface of the wound. We intentionally only assessed the photographic data as this will be the data available from remote patient monitoring.

The data set is heavily imbalanced concerning the representation of the three classes, specifically, the *severe* label is assigned to only 5.1% of all photographs. The distribution for each class is listed in Table 1. The number of photographs per unique patient varies between 1 and 38 with an average of 6.8 photographs. A severe infection case occurred in 17 patients. For these patients on average 2.2 photographs were assigned the severe label.

Table 1. Infection class distribution in the analysed photographic data set

Class	# samples	percentage
No infection	483	66.0 %
Mild infection	212	29.0 %
Severe infection	37	5.1 %
Total	732	100.0 %

The processing steps used in the machine learning classification training procedure were as follows.

1. Detection and filtering of out-of-focus photographs,
2. driveline tube segmentation,
3. prediction of region of interest,
4. classification of wound infection class.

In the following sections each of the processing steps is explained in more detail.

2.1. Detection and filtering of out-of-focus photographs

We would like to filter out highly out-of-focus data samples from the training set to increase the quality of the training data. The aim was to automatically remove the subset of photographs without sufficient detail to determine the infection class.

Quantification of blur in a photograph can be done by computing the sum of the partial second derivatives of the image in both dimensions, known as the Laplacian operator, which has an application in autofocusing for microscopes [2]. The amount of blur is reduced to a single number by taking the variance of the Laplacian value across all pixels in the image.

Before the out-of-focus detection algorithm was developed a set of 692 photographs was available, which were manually classified as either out-of-focus or clear. This allowed us to set a threshold on the variance of the Laplacian that ensures a balanced ratio between precision and recall for out-of-focus detection.

Whenever a device is used to take and send a photo, the out-of-focus detection could trigger an immediate request for a repeated photograph, sent back to the patient's LVAD App while they are still busy with the wound dressing renewal.

2.2. Driveline tube segmentation

Drivelines may have different visual features, from an opaque white colour to transparent, granting view on different internal cable colours, sometimes reflecting flash lighting on their surface. They occur in all photographs and their presence may increase the complexity of training an infection classification network if the network itself is not able to ignore the irrelevant tube features. This section focuses on two separate

approaches for detecting the driveline tubes which allowed us to mask and negate the features in the driveline tube area of the image during infection classification.

In the absence of annotated photographs, a first approach made use of the Felzenszwalb *unsupervised segmentation* algorithm [3]. It is a greedy graph-based algorithm which iteratively merges adjacent pixel regions based on local and global contrast. The Felzenszwalb algorithm is sensitive to the variations within the photographic data, requiring parameter tuning on a per sample basis for adequate segmentation performance, which is inconvenient for practical applications.

A *supervised deep learning* method may be better suitable for capturing the image complexity. In order to facilitate supervised learning we set up a web-based annotation service. Anonymous images were offered to annotators in a random sequence. Images and annotation results were exchanged through a secure connection. LVAD experts were able to use this service to visually annotate driveline regions and other skin coverage (e.g. wound dressing) in photographs. A magnification tool allowed for the exact drawing of the segmentation map with usual point and click devices.

A specific architecture of convolutional neural networks (CNN) called U-net [4] was used for training on the annotated data. It is a type of semantic segmentation CNN which can be used to assign a class label ('driveline tube' or 'background' in this case) to each pixel in an image. Physicians used the annotation service to annotate 185 photographs which we randomly split into 148 training and 37 validation samples. Data augmentation is applied to the training set and ground truth annotations in the form of affine transformations to artificially enrich the training set.

2.3. Prediction of region of interest

Experiments with multiple classification preprocessing configurations showed that selecting a rectangular area around the driveline exit site increases performance compared to using the full image as input to the classification module. This is also due to the wide variety of zooming at the exit sites and wound areas in the data set.

A training/validation set was created by manually annotating 745 photographs. Similar to tube segmentation (Section 2.2) we trained a U-net on this training set to convert image input to region of interest "blobs" as output. The blobs, which indicate a region of interest prediction, were converted to rectangular sections using post-processing, which is a requirement for our classification model.

2.4. Classification of wound infection class

While the first three steps above provide methods and tools for the preparation of the photographs to be analysed, infection class recognition is the main contribution of the research described in this paper. We set up a classification network that learns to identify one of the three infection classes (*none, mild, severe*) based on an input image.

Experiments were set up using a variety of popular CNN classification architectures. The best performing network on our data set was the VGG-16 architecture [5], pretrained on ImageNet [6] and fine-tuned on the driveline photographic data. The training data was augmented using affine transformations to indirectly increase the effectiveness of the classifier [13].

Since the labels of our training set were initially assigned using more information than only the visual features observed in the photographs, we initiated a blind expert evaluation. In such an evaluation we can not only compare the performance of the

classification CNN with respect to the original labels, but also to the performance of human experts in an identical task. The blind evaluation set consisted of 100 photographs, containing an even division of samples from both heart clinics. The chosen class distribution reflects the class distribution of the full data set as much as possible, while ensuring a minimum of 15 samples per class (see Table 3). The images were drawn randomly from the respective class's image pool. Physicians from both clinics were asked to provide their classification of infection class for each of the evaluation photographs. The classification CNN was trained using the leave-one-out method to obtain a single infection class prediction for each of the 100 photographs.

A separate experiment was set up to analyse the effects of tube segmentation on classification performance. Segmentation masks from Section 2.2 were applied to the photographs before feeding these to the classification CNN.

3. Results

3.1. Tube segmentation

The Felzenszwalb *unsupervised* segmentation algorithm was compared to the *supervised* U-net semantic segmentation CNN. We measured the performance of both methods on the annotated validation set (n=37) using the Dice score coefficient. The Dice score quantifies pixel overlap between ground truth annotations and masks generated by segmentation algorithms and is measured in the range of 0 – 1, where 0 is fully dissimilar and 1 is perfect similarity. On the validation set the Felzenszwalb algorithm resulted in a Dice score coefficient of 0.72, while U-net scored higher with a Dice score coefficient of 0.95. An example of Felzenszwalb and U-net output can be seen in Figure 1. In this example the Felzenszwalb algorithm predicted the area around the driveline exit as being part of the driveline, thus masking part of the wound area.

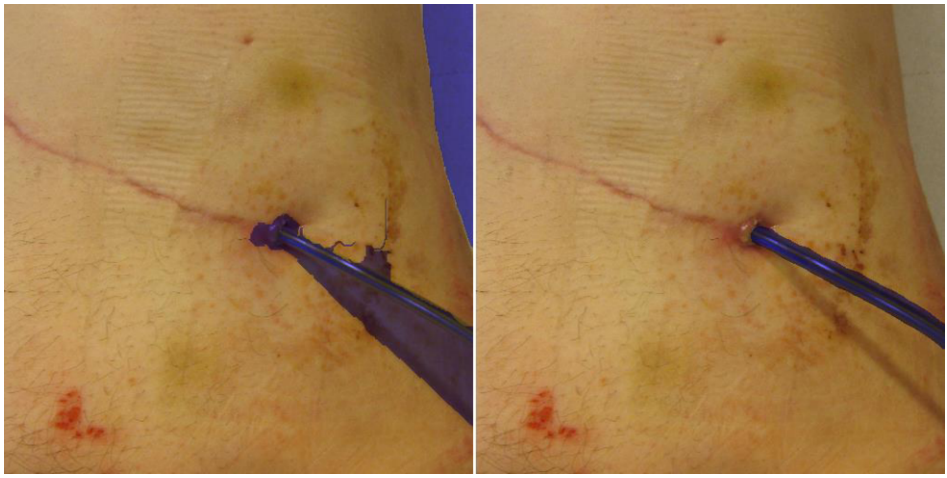

Figure 1. Visualisation of driveline tube segmentation masks. The blue region represents the predicted driveline tube area. Left: Felzenszwalb segmentation method (note that the non-skin background is included in the blue region). Right: U-net segmentation method.

3.2. Prediction of the region of interest

To assess the effect of extracting a region of interest (RoI) on classification performance we compared the classification accuracy on full images, manual RoIs and U-net generated RoIs on a validation set. Table 2 shows the results of each configuration. We observe that cropping RoIs either manually or generated by U-net performed slightly better than using the full image when evaluating infection classification performance. Manually cropped RoIs led to slightly better results than U-net RoIs.

Table 2. Infection classification accuracy and macro (unweighted) F1 score based on different types of region of interest (RoI) extraction methods.

RoI type	Accuracy (%)	F1 score
None (full image)	66.7	0.472
Manual RoI	71.7	0.498
U-net RoI	69.8	0.496

3.3 Infection classification

Two LVAD experts, one from each of the two clinics involved in the study, have assigned infection class labels to each of the 100 photographs in the blind evaluation. For comparison, output predictions from the classification CNN have been obtained on the same set. We compute prediction accuracy using the original labels, and the resulting metrics are shown in Table 3. The total accuracy of all participants, humans and machine, is between 66% and 69%. The mean accuracy is derived from the results of the three classes, weighted by the class distribution, as shown in Table 3. The severe infection class, which is least represented in the data and prone to under-skin processes, shows the lowest accuracy for all participants. Since prediction performance in this class is at least as important as in the other classes, the macro F1 score is reported for each participant as well. We observed that due to the lower performance on the severe infection class the macro F1 average score of the classification CNN is lower than that of the trained physicians.

Multiple approaches for applying tube masks (generated by the U-net segmentation CNN) to classification input photographs were explored, such as setting the driveline tube to a solid colour and a combination of inpainting and blurring to attempt to hide the tubes in the photographs. In every approach in which a tube segmentation mask was applied to classification input images, the resulting classification accuracy ended up lower than without applying the mask.

Table 3. Prediction accuracy and F1 score of each participant providing predictions on the blind evaluation set (n=100). The macro F1 score average is reported for each candidate, which is calculated by weighing each class equally. Numbers in bold indicate the highest scores per class.

	Physician 1		**Physician 2**		**Classification CNN**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
No infection (n=58)	**89.7**	**0.85**	81.0	0.80	81.0	0.80
Mild infection (n=27)	44.4	0.50	51.9	0.47	**66.7**	**0.57**
Severe infection (n=15)	**33.3**	0.34	**33.3**	**0.42**	13.3	0.20
Total / macro (n=100)	**69.0**	**0.56**	66.0	**0.56**	67.0	0.52

4. Discussion

In this study we explored machine learning approaches towards providing assistance in applying photographic data for remote LVAD patient monitoring. Patients or caregivers can provide photographs without needing clinical assistance by using a mobile device such as a smartphone, eventually in combination with a dedicated App that might transmit more data, e.g. from the LVAD device itself.

We have demonstrated that a U-net architecture can achieve high driveline tube segmentation performance (Dice score coefficient of 0.95) with only 148 training samples. When applying the segmentation masks during infection classification we observe that applying the segmentation masks to the input does not improve classification accuracy. It is possible that our attempts to hide the driveline tube distorts the image, affecting relevant features. Alternatively, the network may have learnt to use the tube for more detailed localization of the wound area. Future work could use segmentation masks to further improve region of interest prediction, as one can derive the driveline exit site, and therefore region of interest, from an accurate driveline segmentation mask (for a visual example see Figure 1).

In typical specialized machine learning applications accuracy figures of 90% or higher on multi-class problems are common. In contrast, our classification CNN achieves 67% accuracy on the blind evaluation experiment with three classes. However, we observe that human experts did not score significantly higher on a pure visual infection recognition task.

We can conclude that photographic data is not in all cases sufficient to accurately determine the infection class without additional external data. Future planned developments include the application of thermographic imaging. This is expected to lead to improved results as infrared images can uncover the sub-surface heat sources that are present in infection processes. This development would thus mainly improve the infection classification performance, as more severe infections show an increase in temperature around the wound. Smartphone devices that allow users to simultaneously take a picture in both visible light and infrared light have recently become available on the market.

5. Acknowledgements

The authors of this paper would like to thank Dr. Ioannis Giotis for his valuable knowledge on skin lesion segmentation during the start of the project. In addition, we thank Dr. Rolf Neubert for the coordination between involved parties as well as helping to improve the writing style of the paper.

References

[1] Specifications Manual for Joint Commission National Quality Measures, New York Heart Association (NYHA) Classification, https://manual.jointcommission.org/releases/TJC2016A/DataElem0439.html, last access: 12.02.2019.

[2] J. L. Pech-Pacheco, et al., *Diatom autofocusing in brightfield microscopy: a comparative study,* in: Proceedings. 15th International Conference on Pattern Recognition. IEEE, 2000. pp. 314-317.

[3] P.F. Felzenszwalb, D. P. Huttenlocher, Efficient graph-based image segmentation*, International journal of computer vision*, **59,** 2004, 167-181.

[4] O. Ronneberger, P. Fischer, T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, in: International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015. pp. 234-241.
[5] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.
[6] Stanford University, Princeton University, ImageNet, http://www.image-net.org/, last access: 22.1.2019.
[7] Pinney, S. P., Anyanwu, A. C., Lala, A., Teuteberg, J. J., Uriel, N., & Mehra, M. R. (2017). Left ventricular assist devices for lifelong support. *Journal of the American College of Cardiology*, *69*(23), 2845-2861.
[8] Zierer, A., Melby, S. J., Voeller, R. K., Guthrie, T. J., Ewald, G. A., Shelton, K., ... & Moazami, N. (2007). Late-onset driveline infections: the Achilles' heel of prolonged left ventricular assist device support. *The Annals of thoracic surgery*, *84*(2), 515-520.
[9] Deniz, E., Feldmann, C., Schmidt, T., Hoffmann, J. D., Hanke, J., Rojas-Hernandez, S. V., ... & Haverich, A. (2017). The Impact of Telemonitoring in Patients with Ventricular Assist Device. *The Thoracic and Cardiovascular Surgeon*, *65*(S 01), ePP17.
[10] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115.
[11] Masood, A., & Ali Al-Jumaily, A. (2013). Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. *International journal of biomedical imaging*, *2013*.
[12] Wernick, M. N., Yang, Y., Brankov, J. G., Yourganov, G., & Strother, S. C. (2010). Machine learning in medical imaging. *IEEE signal processing magazine*, *27*(4), 25-38.
[13] Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.