XNet: A convolutional neural network (CNN) implementation for medical X-Ray image segmentation suitable for small datasets

Joseph Bullock, Carolina Cuesta-Lázaro, and Arnau Quera-Bofarull

Department of Physics, Durham University, UK

ABSTRACT

X-Ray image enhancement, along with many other medical image processing applications, requires the segmentation of images into bone, soft tissue, and open beam regions. We apply a machine learning approach to this problem, presenting an end-to-end solution which results in robust and efficient inference. Since medical institutions frequently do not have the resources to process and label the large quantity of X-Ray images usually needed for neural network training, we design an end-to-end solution for small datasets, while achieving state-of-the-art results. Our implementation produces an overall accuracy of 92%, F1 score of 0.92, and an AUC of 0.98, surpassing classical image processing techniques, such as clustering and entropy based methods, while improving upon the output of existing neural networks used for segmentation in non-medical contexts. The code used for this project is available online.[1]

Keywords: X-Ray image segmentation, medical image processing, machine learning.

1. INTRODUCTION

X-Ray image segmentation is of great importance in many medical applications such as image enhancement, and other processing tasks, computer assisted surgery, and anomaly detection. These applications regularly require the segmentation of images into 3 categories: open beam, soft tissue and bone. Current methods rely heavily on a complex system of classical image processing techniques, such as clustering approaches,[2,3] line fluctuation analysis,[4] or entropy-based methods,[5,6] which often require the tuning of hyperparameters for each body part class. However, utilising machine learning offers several advantages over these traditional methods since: (i) it naturally addresses noise, (ii) it generalises well to different body parts, and (iii) the segmented regions have continuous boundaries.

Although the use of machine learning in the healthcare sector has grown significantly in recent years, and with it open source datasets have become more readily available,[7–10] we know of no publicly available labelled dataset of X-Ray images that can be used in training a neural network for segmentation tasks as presented in this paper. This means institutions wishing to train such a network must provide and label their own images. X-Ray images are expensive to obtain, and manual labelling is time consuming, thus acquiring a large and varied database may not be possible; a common issue raised when discussing the feasibility of neural networks for X-Ray segmentation.[11]

We design a unique Convolutional Neural Network (CNN) architecture to perform segmentation by extracting fine grained features, while controlling the number of trainable parameters to prevent overfitting. The network is trained on 150 X-Ray images, with no scatter correction and comprising of 19 body parts in an imbalanced way. This dataset, we believe, is of a manageable size to be created by a medical institution. We present our network, and a full end-to-end description of it's implementation, including post-processing stages for the minimisation of false positives, and optimisation of the F1 score. Despite our dataset being small compared to those used in many machine learning applications, we achieve an overall accuracy significantly higher, and more generalisable, than

Further author information: (Send correspondence to Joseph Bullock)
Joseph Bullock: E-mail: j.p.bullock[at]durham.ac.uk,
Carolina Cuesta-Lázaro: E-mail: carolina.cuesta-lazaro[at]durham.ac.uk,
Arnau Quera-Bofarull: E-mail: arnau.quera-bofarull[at]durham.ac.uk

work using classical image processing techniques.[2–6,12] Additionally, we show that our architecture outperforms leading image segmentation networks developed for other applications.[13]

Our paper is structured as follows: after reviewing a selection of the existing literature in Section 2, we discuss our dataset in Section 3 - how we collect and label the data, and the augmentation methods to prevent overfitting; in Section 4 we address the design and structure of our CNN, covering training and testing stages; the results of the network are discussed in Section 5, where we also address the post-processing stage of false positive reduction; a comparison of our results with other works from the classical and machine learning literature is presented in Section 6, after which we discuss future applications and developments in Section 7.

2. RELATED WORK

Note that we provide this section not as an extensive literature review, but to lay the groundwork for where the methodology and results of this work sit in the current research landscape. We include work from which we have drawn important insights and information, while attempting to provide a pedagogical introduction to the development of the field of X-Ray image segmentation, along with modern image segmentation methods in a broader context.

Methods for segmenting X-Ray images have been a constant topic in the literature for many years, due to its role in image processing and other analysis based operations. Much previous work focuses on using classical image processing techniques.[11,14] Pixel clustering based on similarity in certain parameters is a commonly employed technique. Kubilay Pakin *et al.*[2] use clustering as a component of their segmentation algorithm, achieving high accuracy scores, but requiring hyperparameter tuning for each body part class, while struggling to produce smooth boundaries. Similarly, good results have been achieved by Wu and Mahfouz,[3] who employ spectral clustering methods and produce much smoother boundaries, however, this application was tuned purely to knee image analysis. Entropy-based approaches have been employed by Bandyopadhyay *et al.*,[5,6] which have enabled clean boundary identification, yet suffer from extraneous edges such as bone cracks or image distortions. Kazeminia *et al.*[4] build on this work, employing existing edge detection algorithms such as *Sobel*[15] and *Canny*,[16] while analysing the intensity fluctuations in pixel rows to more accurately select the bone boundary. This work is less sensitive to noise, however, loses some boundary continuity. Similarly, atlas models have been developed for medical image segmentation, and an application to rib cage segmentation is given by Candemir *et al.*[12] From such techniques the authors are able to generate complete segmentations, however, the boundaries remain noisy and the area under the ROC curve (AUC) is highly dependent on the dataset analysed.

The growth of machine learning applications in the healthcare sector has been considerable in recent years. Indeed, due to the large amount of information encoded in X-Ray images, focused research into their analysis has been significant. Aiding this advancement of research has been an increase in availability of X-Ray image datasets, each tailored to different applications.[7–10] Using such datasets, many machine learning approaches for the detection of anomalies, such as pneumonia,[17] pulmonary tuberculosis,[18] and thoracic diseased[19] have been developed, as well as *diagnosing* a variety of diseases based on chest X-Rays.[20] Islam *et al.*[21] and Qin *et al.*[22] provide comparisons of a range of neural networks applied to detecting anomalies in chest X-Rays. These detection technologies usually perform image segmentation to localise the position of the anomaly, or of a certain bone structure (most commonly the rib cage), and are carefully tuned to these applications. Additionally, due to their architectures, many of the networks presented in the above literature would not be suitable for performing a pixel-level segmentation over the entirety of the image.

Complete multi-class image segmentation is an active, high-growth area of research, most recently driven by autonomous vehicle development. Neural networks have been at the forefront of this research and take various forms including the encoder-decoder design similar to that presented in this paper,[23–27] and fully connected networks.[28] Several applications of these networks also utilise the technique of image augmentation to aid network generalisation and reduce the risk of overfitting. Similarly, Badrinarayanan, Handa and Cipolla[13] present a simplified version of the widely applied SegNet architecture,[23] showing improved performance on small datasets. Indeed, it is against this simplified network, known as **SegNet-Basic**,[13] that we benchmark our network performance. It is noted that some networks have been specifically designed for the total segmentation of medical images, yet these applications have been largely constrained to the segmentation of cell structures.[27,29,30]

There have been examples of networks, such as U-Net,[27] being applied to other segmentation tasks in the field of X-Ray image segmentation,[31] but only in specific use cases.

3. DATA

3.1 Collection and Labelling

We use data collected from two sources at IBEX Innovations Ltd.:[32] 69 CT scans images of feet, knees and phantom heads, and 81 standard X-Ray images of different body and phantom body parts, of which the thorax is the most underrepresented. Several images contain foreign metal objects, such screws and staples, which we aim to classify as bone. The particular distribution among body part classes can be seen in Table 1. The images have not been corrected for scatter effects; but they have been dark-corrected by removing the detector background signal. Additionally, we pre-process each image by performing mean subtraction and pixel value normalisation to within the range $[-1, 1]$. The images are labelled using the free software Gimp,[33] by assigning a colour to each of the three distinct regions: open beam, soft tissue and bone. As a case study we specifically aim to minimise the soft tissue false positives, and so primarily avoid labelling any bone or open beam region as soft tissue. Both the images and labels are resized to 200 × 200 pixels to match the network input shape.

Bodypart list			
Ankle	10	Leg	1
Arm	3	Lumbar spine	6
Cervical	1	Neck of femur	15
Chest	1	Pelvis	2
Elbow	1	Shoulder	2
Femur	3	Thigh	8
Foot	29	Thorax	1
Hand	4	Tibia	4
Head	11	Wrist	11
Knee	36	**Total:**	**150**

Table 1: Number of images in each body part class on our dataset. Note that the different body part classes are highly unbalanced.

3.2 Augmentation

Compared to previous successful applications of neural networks to image segmentation, our dataset is small. There are 150 images in total, with unbalanced body part classes, representative of that which may be created by a medical institution. Therefore, we artificially augment the training images with the two-fold purpose of creating a larger dataset, to avoid overfitting, and balancing the different body part classes through augmented oversampling. After experimenting with a variety of filters, we find that elastic transformations are a crucial component for generating realistic augmented images, combined with translations, rotations, shear and cropping. For our dataset, we also find augmenting to produce 500 images per body part class gives the highest validation accuracy. If there is only one example of a particular body part, this image is placed in the test set and not used for training.

4. ARCHITECTURE

4.1 XNet

XNet is based on an *encoder - decoder* style architecture commonly used in image segmentation.[23–27]

Encoder The encoder consists of a series of convolutional layers, for feature extraction, and max pooling layers to downsample the input image. Breaking up the downsampling into a multiple stages allows for varying levels of extraction, with increasingly global features learnt through the convolutional layers at each pooling stage.

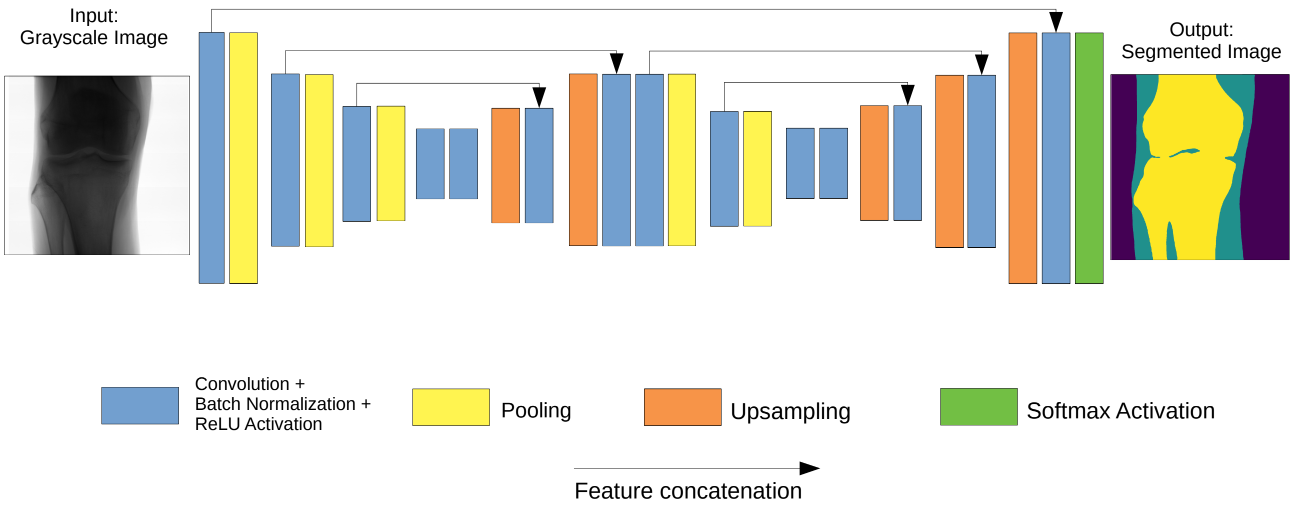


Figure 1: Visualisation of XNet architecture including example input image, left, and output segmented mask, right. Feature concatenation of same dimension layers helps to avoid losing fine-grained detail. Softmax activation function provides final pixel-wise classification.

Decoder After feature extraction, the decoder performs upsampling to generate a segmented mask of equal dimension to the input image. Similar to the encoder, using a multistage upsampling process with convolutional layers in between allows for varying degrees of fine grain feature reconstruction during upsampling, thus producing dense feature maps.

Due to the small size of our dataset, we avoid large *serial* downsampling of the input image compared to many other networks, particularly those used in image classification. We avoid this since performing a greater number of downsamplings in series can be detrimental to accurate boundary level detail, particularly around smaller structures. However, downsampling allows for learnable feature extraction, and so is important to include in the network.

We present an architecture which incorporates a comparable, or greater, number of downsampling stages for feature extraction as other segmentation networks, whilst avoiding overly-reducing image resolution. This is achieved by using two encoder-decoder modules in succession, whilst storing encoder feature maps and using them during the creation of the dense feature maps in the decoders (as can be seen in Figure 1).

Each convolutional layer is associated with a rectified-linear non-linearity (ReLU) activation function, and the decoder upsamples the image using nearest-neighbours upsampling. Storage and use of the encoder feature maps is performed through filter copying between layers of equivalent dimensions, meaning the model is less likely to 'forget' what it has previously learnt, thus decreasing the likelihood of loosing fine-grained detail after performing downsampling (a technique used in several encoder-decoder networks[23,27]). At each convolutional layer we employ L2 norm regularisation, with penalty parameter $\lambda = 5 \times 10^{-4}$, to improve network generalisation and prevent overfitting.[34] For more detailed information regarding network architecture and parameters, including filter sizes, see the supplementary material.[1]

4.2 Training

We train our model on augmented data, thus increasing the number of training examples, thereby reducing the chance of overfitting. Of the 150 images in our original dataset, 108 are set aside for augmentation and training. The remaining images are used in validation and testing stages in equal proportion.

The ground truth masks used for training are one-hot encoded. For each input pixel X we optimise the categorical cross-entropy loss,

$$L(X,y) = -\sum_i I(y,i) \log p(Y=i|X), \tag{1}$$

where y is the output label generated by the network, $p(Y=i|X)$ is the probability that the network assigns the label i given the input data, and $I(y,i)$ is the indicator function defined by

$$I(y,i) = \begin{cases} 0 & \text{if } y \neq i, \\ 1 & \text{if } y = i. \end{cases} \tag{2}$$

We train using Adam optimisation[35] with learning rate 10^{-4}.

Since our aim is to design a network that could also be retrained by a non-specialist institution who may not have access to large memory GPUs, we choose small mini-batch and kernel sizes. We train on a batch size of 5 with each convolutional layer having a kernel size of 3×3. To optimise training time we use *Early Stopping* (see Section 7.8 of Goodfellow, Bengio and Courville[36]) by monitoring 'validation loss' with a patience of 20. The validation set is not augmented and chosen by randomly selected at least one image from each body part class, so as to avoid tuning to a bias dataset. Training using the above parameters took 7 hours on a GTX 1060 6GB GPU.

4.3 Testing

We test our network on images without augmentation chosen in a similar way to the validation set. Indeed, we manually classify certain images as 'difficult' based on factors such as: bone structure complexity, noise, and the contrast ratio. To ensure our network can handle such difficulty, and generalise well, we ensure that the test set contained a disproportionate number of these difficult cases. We carry out all testing before post-processing, and the accuracy score is calculated using the *categorical accuracy* metric in Keras.[37]

Although we correct for body part class imbalance at the pre-processing stage, our dataset is still highly imbalanced in the segmentation categories. The open beam region is both the most prevalent in the dataset and the easiest to classify (during architecture development and hyperparameter tuning this category always achieved the highest accuracy). Therefore, accuracy over all classes is not necessarily the most effective metric by which to measure the model's performance.

For our applications we look for a balance between false positive reduction and true positive enhancement. Therefore we use the F1 score as a measure of network performance. The comparison of accuracy and F1 scores can be found in Section 5.

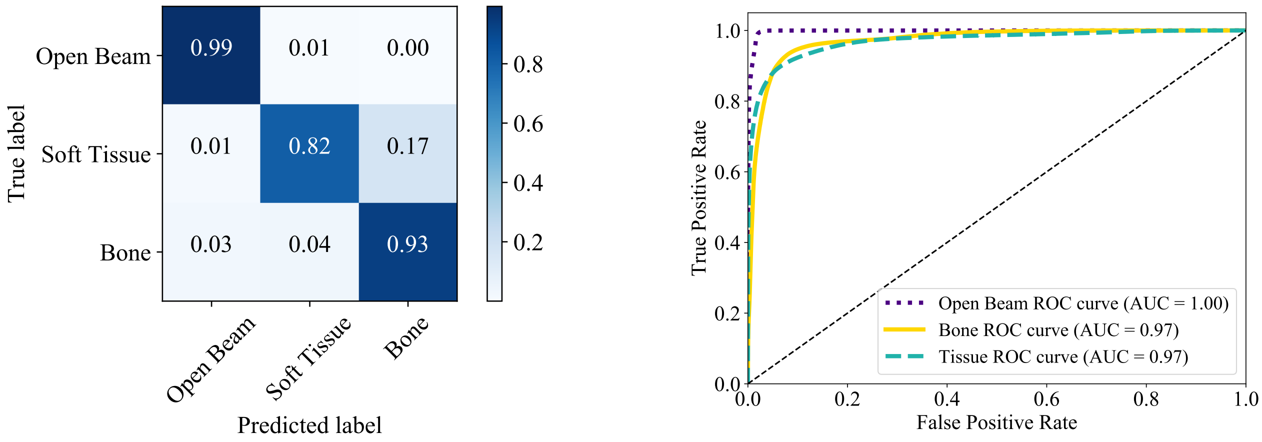
5. RESULTS

5.1 Network Performance

The network achieves an open beam, soft tissue, and bone classification accuracy of 96%, 94%, and 88% respectively, and an overall weighted averaged accuracy of 92%. Additionally, we obtain an F1 score of 0.97, 0.87, and 0.90, in the open beam, soft tissue, and bone categories respectively. Taking into account class imbalance, the overall weight averaged F1 score is 0.92. The summary of results is presented in Table 2. In Figure 3, we show some predictions of images in our test set, with the confusion matrix computed over the test set presented in Figure 2a. The network can be seen to perform well for different body parts, including on the more challenging, due to its complexity and absence of training examples, chest region.

Category	F1-Score	AUC	Accuracy	Confidence
Open Beam	0.97	1.00	96%	99%
Soft tissue	0.87	0.97	94%	95%
Bone	0.90	0.97	88%	97%
Weighted average	**0.92**	**0.98**	**92%**	**97%**

Table 2: Evaluation metrics for the three categories, and their weighted averages.



(a) Confusion matrix

(b) ROC curve

Figure 2: (a) Normalised confusion matrix of XNet results. Each row represents the instances in a class, while each column shows the class predicted by the network for those instances. Here, most of the errors are made classifying soft tissue as bone. (b) ROC curve showing the true positive rate against the false positive rate for the open-beam, bone and soft-tissue categories. The area under the curve (AUC) for the different categories is shown.

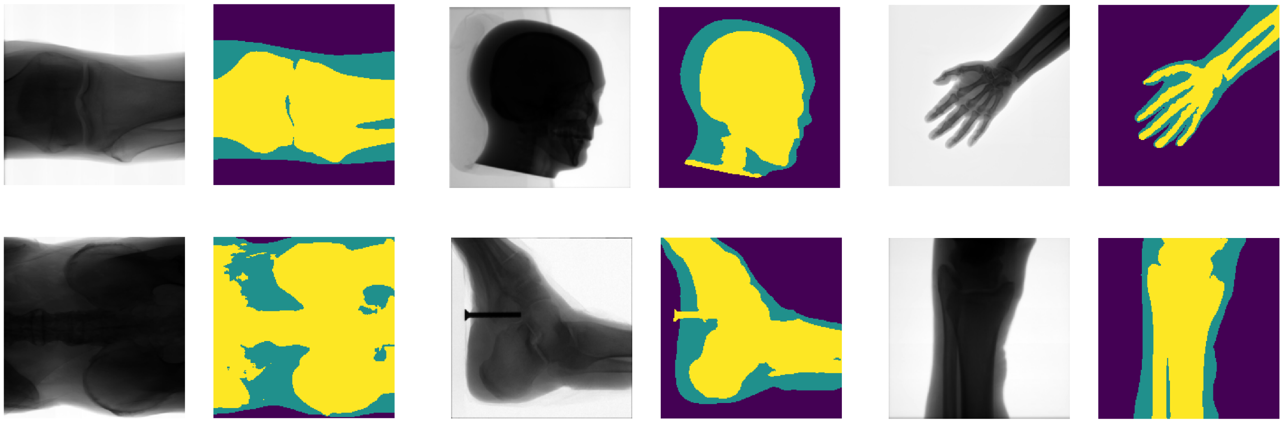


Figure 3: Segmentation predictions from the test set. Top row images show a knee, phantom head and hand. Bottom row shows a pelvis, ankle with a metallic bolt, and the lower half of a leg. The open beam area is shown in purple. Bone is shown in yellow and soft tissue in green.

5.2 Calibration

Modern CNNs are often ill-calibrated, making them overconfident about their predictions.[38] For networks broadly applicable to a variety of medical image segmentation tasks, such calibration error can be detrimental to important post-processing steps, such as false positive reduction in a given category. The output of XNet is a 3-dimensional probability map, where each pixel is assigned a probability of belonging to one of the different categories. We define network confidence in a given category as

$$\mathrm{conf}(X) = \sum_{i \in X} p_i, \tag{3}$$

where X is the set of all pixels assigned to that category, and p_i is the probability that the ith pixel belongs to said category.

A well calibrated network is defined as a network whose confidence averaged over all output categories, is close to the network's averaged accuracy.[38] We find that our network is indeed well calibrated, as defined by this metric, as can be seen in Figure 4. XNet is therefore well suited for the employment of a variety of post-processing techniques to tailor the output depending on the use case.

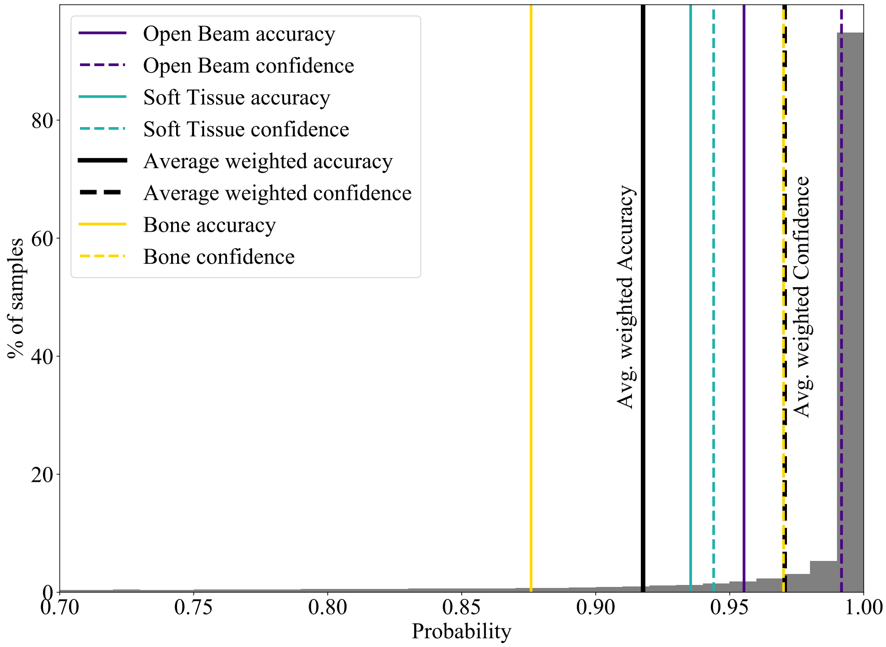


Figure 4: Confidence histogram across all categories, showing that the network is well calibrated. For the individual categories, the best calibrated is soft tissue with a difference of 1%. For open beam the difference is 3%. The least calibrated category is bone, with a difference of 9%. Note: as mentioned, in our case study, when we carry out hyperparamter optimisation we are focusing on minimising soft tissue false positives, so these results are as expected.

5.3 False Positive Reduction

The simplest way of obtaining a segmentation from the network output is to assign each pixel to the most likely category. Nonetheless, with our objective of reducing soft tissue false positives in the soft tissue category, we classify a pixel as soft tissue only if the network predicted probability is higher than a given threshold. Since our network is well calibrated, this probability threshold reflects the confidence that we demand for a pixel to be in this category. We find that increasing this probability threshold reduces the number of false positives at the expense of the number of true positives. For example, with a probability threshold of 0.90, the false positive rate in the soft tissue class is reduced to 3%, while reducing the true positive rate to 71%. Choosing the ideal probability threshold depends on the particular use case. We choose more severe thresholds for the body parts that the network struggles more to segment. We obtain an area under the curve of 1.00 for the open beam region and 0.97 for both the bone and soft tissue classes, thus demonstrating the high classification ability of our network.

6. CONCLUSION

We develop a fully automatic method to segment medical X-Ray images given a small dataset. As a compromise between having a deep network to extract high-level features and fine-grained detail, and a network that avoids overfitting to small datasets, we present an architecture with two encoder-decoder modules. We train this network on a dataset consisting of 150 images, artificially augmented to generate 7000 training images. Evaluating the network performance we find an overall accuracy of 92% and F1 score of 0.92, with an AUC of 0.98.

We benchmark our network against the popular SegNet design. Starting with the Segnet-Basic architecture,[13] and carrying out a hyperparameter search similar to the one we do for XNet, the best result obtained gives a 2% improvement on bone classification accuracy with respect to XNet, but at the expense of only having a 75% true positive rate and 25% false positive rate in the soft tissue category (in comparison to our network which achieves an 82% TP ratio, see Figure 2a). Additionally, we significantly outperform this network when detecting the open beam region. SegNet-Basic achieves F1 scores of 0.96, 0.83, and 0.90 for open beam, soft tissue and bone regions respectively, with an overall weighted F1 score of 0.89, thus giving a lower F1 score in every category compared to XNet (see Table 2). The full implementation of SegNet[23] has too many parameters to effectively fit to such a small dataset. Similarly, its size requires significantly more computational power, thus making it impractical for the purpose of being trainable by many medical institutions.

We also see a significant improvement when benchmarking against state-of-the-art classical image processing techniques. Figure 5 shows a comparison between our architecture and the work of Kazeminia *et al.*,[4] who built upon that of Bandyopadhyay *et al.*[5,6] The XNet segmentation produces smoothly connected boundaries around the bone regions, in addition to differentiating well between bone and soft tissue regions. It should be noted that in producing this output, our algorithm was trained on a set of high resolution TIF images, as produced by the X-Ray scanner, whereas this analysis was run on significantly lower quality JPEG images, meaning our network could not achieve its full potential on this image. Additionally, our training set does not contain any feet viewed from the angle as seen in Figure 5, thus demonstrating the generalisability of our network.

Another classical example is the clustering based method presented by Kubilay Pakin *et al.*,[2] which gives a 92% accuracy averaged over 14 images belonging to 5 body part classes. This accuracy score is calculated after each of the two free parameters of the algorithm are fine-tuned for each individual image. Our approach obtains a similar accuracy score in a larger, more diverse dataset, generalising well to 19 body part classes without specific tuning.

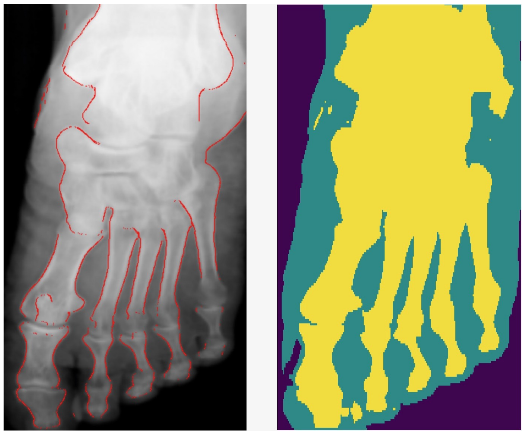


Figure 5: Comparison of our algorithm, right, with the work of Kazeminia *et al.*,[4] left, showing significant improvement in boundary smoothness

7. FUTURE WORK

As an improvement to our implementation, we wish to focus on reducing the false positive rate at the post-processing stage. Additionally, there have been several attempts at refining segmentation mask outputs, such as using conditional random fields.[26] Such methods could be applied to our work to further improve boundary smoothness, and reduce the likelihood of false islands appearing in the mask. Another approach could be to train an adversarial network, along with XNet, to detect inconsistencies between the network generated segmented maps and the ground truth. However, such an implementation would significantly increase the training time of the network, while also increasing the risk of falling into local minima during training.

Furthermore, it would be interesting to diversify the classification categories to identify different bone types or tissue materials. Identification of such differences, particularly in the tissue region, is of great importance to the medical field since tissue abnormality detection is a highly non-trivial task when using X-Ray images.

8. ACKNOWLEDGEMENTS

The authors are part of the Centre for Doctoral Training in Data Intensive Science at Durham University supported by STFC grant number ST/P006744/1. The authors are grateful to IBEX Innovations Ltd.[32] for hosting them and providing data, and to Kazeminia *et al.*[4] and Bandyopadhyay *et al.*[6] for providing images used in their work. This work made use of the facilities of the Hamilton HPC Service of Durham University. This work also used the DiRAC Data Centric system at Durham University, operated by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). This equipment was funded by BIS National E-infrastructure capital grant ST/K00042X/1, STFC capital grant ST/H008519/1, and STFC DiRAC Operations grant ST/K003267/1 and Durham University. DiRAC is part of the National E-Infrastructure. The authors are grateful to the above organisations for their support.

References

[1] Bullock, J., Cuesta-Lázaro, C., and Quera-Bofarull, A., "XNet." https://github.com/JosephPB/XNet (2018).

[2] Kubilay Pakin, S., Gaborski, R. S., Barski, L. L., Foos, D. H., and Parker, K. J., "Clustering approach to bone and soft tissue segmentation of digital radiographic images of extremities," *Journal of Electronic Imaging* **12**, 12 – 12 – 10 (2003).

[3] Wu, J. and Mahfouz, M. R., "Robust x-ray image segmentation by spectral clustering and active shape model," *Journal of Medical Imaging* **3**(3), 034005 (2016).

[4] Kazeminia, S., Karimi, N., Mirmahboub, B., Soroushmehr, S. M. R., Samavi, S., and Najarian, K., "Bone extraction in x-ray images by analysis of line fluctuations," in [*2015 IEEE International Conference on Image Processing (ICIP)*], 882–886 (2015).

[5] Bandyopadhyay, O., Chanda, B., and Bhattacharya, B., "Entropy-based automatic segmentation of bones in digital x-ray images," *Pattern Recognition and Machine Intelligence* , 122–129 (2011).

[6] Bandyopadhyay, O., Biswas, A., Chanda, B., and Bhattacharya, B., "Bone contour tracing in digital x-ray images based on adaptive thresholding," *Pattern Recognition and Machine Intelligence* **852**, 465–473 (2013).

[7] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M., "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in [*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 3462–3471 (2017).

[8] Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R. L., Langlotz, C., Shpanskaya, K., Lungren, M. P., and Ng, A. Y., "MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs," *arXiv preprint arXiv:1712.06957* (2017).

[9] Yan, K., Wang, X., Lu, L., and Summers, R. M., "Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of Medical Imaging* **5**, 5 – 5 – 11 (2018).

[10] Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J., "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association : JAMIA* **23**(2), 304–310 (2016).

[11] Sharma, N. and Aggarwal, L. M., "Automated medical image segmentation techniques," *Journal of Medical Physics / Association of Medical Physicists of India* **35**(1), 3–14 (2010).

[12] Candemir, S., Jaeger, S., Antani, S., Bagci, U., Folio, L. R., Xu, Z., and Thoma, G., "Atlas-based rib-bone detection in chest x-rays," *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* **51**, 32–39 (2016).

[13] Badrinarayanan, V., Handa, A., and Cipolla, R., "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv:1505.07293* (2015).

[14] Mansoor, A., Bagci, U., Foster, B., Xu, Z., Papadakis, G. Z., Folio, L. R., Udupa, J. K., and Mollura, D. J., "Segmentation and image analysis of abnormal lungs at ct: Current approaches, challenges, and future trends," *Radiographics* **35**(4), 1056–1076 (2015).

[15] Sobel, Irwin, F. G., "An isotropic 3x3 image gradient operator," Presentation at Stanford A.I. Project (1968).

[16] Canny, J., "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**(6), 679–698 (1986).

[17] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., and Ng, A. Y., "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225* (2017).

[18] Lakhani, P. and Sundaram, B., "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology* **284**(2), 574–582 (2017). PMID: 28436741.

[19] Yan, C., Yao, J., Li, R., Xu, Z., and Huang, J., "Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays," *arXiv preprint arXiv:1807.06067* (2018).

[20] Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., and Lyman, K., "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv preprint arXiv:1710.10501* (2017).

[21] Islam, M. T., Aowal, M. A., Minhaz, A. T., and Ashraf, K., "Abnormality detection and localization in chest x-rays using deep convolutional neural networks," *arXiv preprint arXiv:1705.09850* (2017).

[22] Qin, C., Yao, D., Shi, Y., and Song, Z., "Computer-aided detection in chest radiography based on artificial intelligence: a survey," *BioMedical Engineering OnLine* **17**, 113 (2018).

[23] Badrinarayanan, V., Kendall, A., and Cipolla, R., "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(12), 2481–2495 (2017).

[24] Hong, S., Noh, H., and Han, B., "Decoupled deep neural network for semi-supervised semantic segmentation," in [*Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*], *NIPS'15*, 1495–1503, MIT Press, Cambridge, MA, USA (2015).

[25] Noh, H., Hong, S., and Han, B., "Learning deconvolution network for semantic segmentation," in [*Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*], *ICCV '15*, 1520–1528, IEEE Computer Society, Washington, DC, USA (2015).

[26] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. S., "Conditional random fields as recurrent neural networks," in [*2015 IEEE International Conference on Computer Vision (ICCV)*], 1529–1537 (2015).

[27] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*], 234–241, Springer International Publishing, Cham (2015).

[28] Long, J., Shelhamer, E., and Darrell, T., "Fully convolutional networks for semantic segmentation," in [*2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 3431–3440 (2015).

[29] Ciresan, D., Giusti, A., Gambardella, L. M., and Schmidhuber, J., "Deep neural networks segment neuronal membranes in electron microscopy images," in [*Advances in Neural Information Processing Systems 25*], Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., eds., 2843–2851, Curran Associates, Inc. (2012).

[30] Arganda-Carreras, I., Turaga, S. C., Berger, D. R., Cire?an, D., Giusti, A., Gambardella, L. M., Schmidhuber, J., Laptev, D., Dwivedi, S., Buhmann, J. M., Liu, T., Seyedhosseini, M., Tasdizen, T., Kamentsky, L., Burget, R., Uher, V., Tan, X., Sun, C., Pham, T. D., Bas, E., Uzunbas, M. G., Cardona, A., Schindelin, J., and Seung, H. S., "Crowdsourcing the creation of image segmentation algorithms for connectomics," *Frontiers in Neuroanatomy* **9**, 142 (2015).

[31] Norman, B., Pedoia, V., and Majumdar, S., "Use of 2d u-net convolutional neural networks for automated cartilage and meniscus segmentation of knee mr imaging data to determine relaxometry and morphometry," *Radiology* **288**(1), 177–185 (2018).

[32] IBEX Innovations Ltd., http://ibexinnovations.co.uk.

[33] "GNU Image Manipulation Program (GIMP)." https://www.gimp.org/. Accessed: 2018.

[34] Krogh, A. and Hertz, J. A., "A simple weight decay can improve generalization," in [*Advances in Neural Information Processing Systems 4*], Moody, J. E., Hanson, S. J., and Lippmann, R. P., eds., 950–957, Morgan-Kaufmann (1992).

[35] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).

[36] Goodfellow, I., Bengio, Y., and Courville, A., [*Deep Learning*], MIT Press (2016). http://www.deeplearningbook.org.

[37] Chollet, F. et al., "Keras." https://keras.io (2015).

[38] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q., "On Calibration of Modern Neural Networks," *arXiv preprint arXiv:1706.04599* (2017).