Manuscript NO: 87733

Title: Automated decision support for Hallux Valgus treatment options using anteroposterior foot radiographs

Manuscript Type: Basic Study

Authors: Konrad Kwolek, Artur Gądek, Kamil Kwolek, Radek Kolecki, Henryk Liszka

Dear Editor,

thank you very much for your remarks as well as the reviewer's comments, which permitted us to improve our work. Please find below the responses to comments addressed point-by-point. Each comment has been answered accordingly in the revised manuscript, where the corresponding author added content or modification that has been done is highlighted in the yellow color. Please find in the attachment the revised manuscript in Word format (file name "87733-Revised Manuscript").

We hope that the revised paper will fulfill the requirements for publication in the World Journal of Orthopedics.

Thank you very much.

Best regards,

Henryk Liszka

Language polishing requirements.

The co-author of the publication (Radek Kolecki), an American by origin, performed further language polishing. A confirmation certificates are attached.

Reply to reviewer's comments:

Comment #1: In the introduction, there is limited coverage of the research work related to Hallux valgus (HV) and its associated studies. It is recommended that the authors provide additional information on relevant and recent research in this area.

Response: Thank you very much for this valuable comment. The Introduction has been improved as follows:

Through the use of WBCT scans, it has been demonstrated that up to 87% of hallux valgus cases exhibit metatarsal bone pronation, emphasizing the intricate multiplanar nature of this deformity. This metatarsal pronation explains the perceived metatarsal bone shape and the misalignment of the medial sesamoid bone in radiological studies, which has been recognized as a significant factor contributing to recurrence following treatment. As a result, distal metatarsal articular angle has proved unreliable, demonstrating a poor interobserver agreement[12, 13]. Further research is needed to develop effective approaches for addressing the rotational deformity in individuals with HV[2, 14, 15].

Comment #2: In Figure 3, the authors provide a brief introduction to the data flow of their proposed method. It is recommended that the authors provide more detailed information on this in order to improve clarity.

Response: We agree with you. We extended the caption of Fig. 3 as follows:

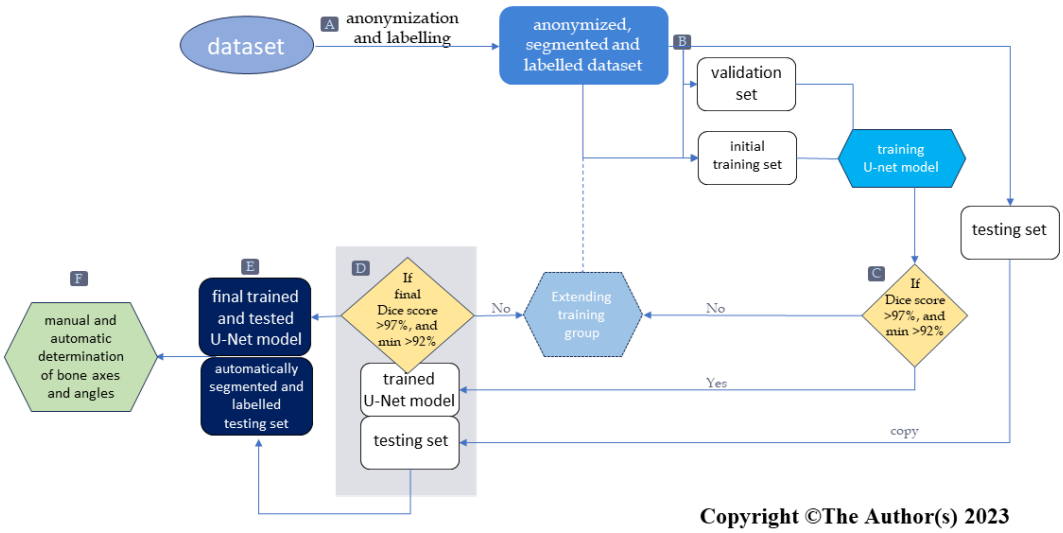


Figure 3 Data flow in the proposed approach. Bones are manually segmented and labelled from anonymized input radiographs to perform multi-class segmentation using a U-Net neural network (A). Radiographs are then randomly assigned to three subsets: training, validation, and testing (B). The accuracy of bone segmentation in each training cycle of the U-Net is validated on a fixed validation subset consisting of 20 radiographs. The U-Net network is trained on a training subset initially consisting of 50 radiographs, which is increased by 10 each training cycle until achieving average SDI > 97% on the validation set (C). Once the network achieves an SDI > 0.97, calculated on the testing subset, the U-Net model completes. If SDI is not > 0.97, the training subset is extended and the U-Net is retrained (D). The final U-Net is used to segment and label bones on all testing radiographs (E). These are used to automatically determine reference points and measure HVA and IMA (F).

Comment #3: In the section of anonymization and manual labelling, it is recommended that the authors provide clarification on whether the data was annotated by medical professionals and how the accuracy of the annotations was ensured.

Response: We are grateful for this remark, which permitted us to improve the paper.

The bones were segmented by a foot surgeon, and checked by another very experienced foot surgeon. To precisely delineate the soft tissue from the bones (to ensure high annotation accuracy) the delineation was performed on high-resolution radiograms, i.e. on anonymized original radiograms. This task was quite time-consuming, but the delineation of bones is precise. Moreover, owing to the modern understanding of pronation and types of shape of the first metatarsal head in hallux valgus deformation described precisely by Wagner et al. [36], the first metatarsal head and the sesamoid bones have been delineated carefully and precisely by a foot surgeon.

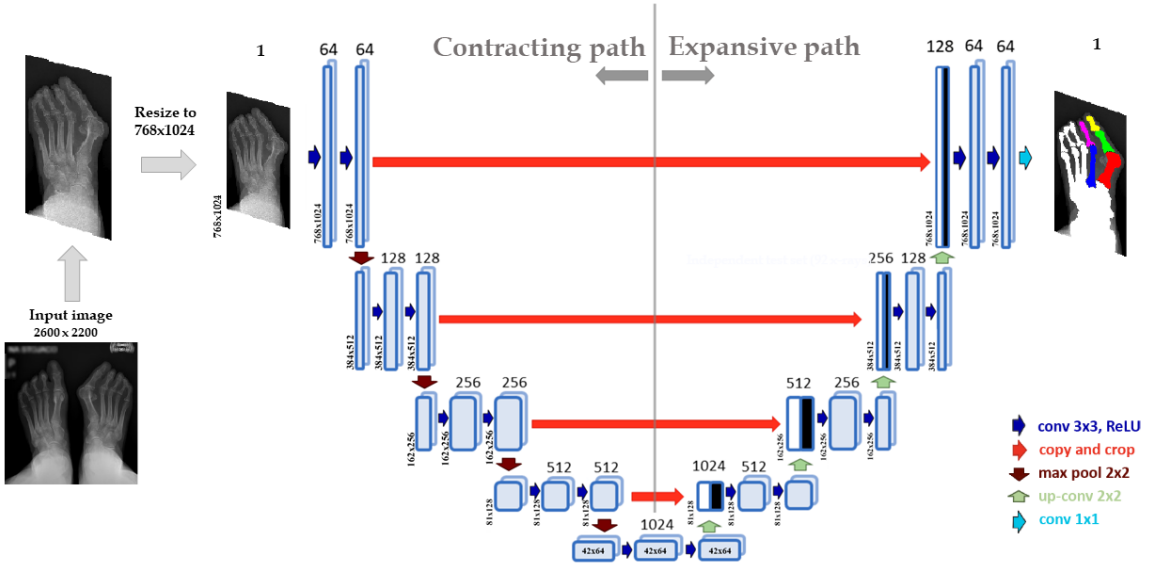
Original text:
In order to train a U-Net network that would achieve high bone segmentation accuracy, bones were manually annotated on original resolution radiographs. (…) To achieve precise measurements of the HVA/IMA, the sesamoid bones must be precisely excluded on manually segmented images[11]. Automated segmentation (delineation) of specific bones from radiographs is particularly challenging due to the complex structure of bones in anteroposterior feet radiographs[36].

Revised text:
To train a U-Net network that would achieve high bone segmentation accuracy, bones were manually annotated on original high-resolution radiographs. (…) Considering the current understanding of pronation and variable shape of the first metatarsal head in hallux valgus deformation described by Wagner et al.[14], the first metatarsal head and the sesamoid bones were delineated carefully and precisely by a foot surgeon to achieve precise measurements of the HVA/IMA [11]. The complex structure of bones in anteroposterior feet radiographs makes automated segmentation (delineation) particularly difficult[38].

14 **Wagner E**, Wagner P. Metatarsal Pronation in Hallux Valgus Deformity: A Review. *J. of the Am. Academy of Orthopaedic Surgeons Global Research and Reviews* 2020; **4**(6)[DOI: 10.5435/JAAOSGlobal-D-20-00091]

Comment #4: In Figure 4, it is recommended that the authors add a legend in the bottom right corner of the U-Net neural network architecture diagram to explain the meaning of the different colored arrows.

Response: Thank you very much for this valuable comment. On the revised figure the arrows depicting the convolution, max pool, up-conv and the copy and crop are in different colors. The meaning of colors is explained in the legend of the figure.



Comment #5: In the section of Radiographs Pre-processing, the authors designed a U-Net neural network for bone segmentation. It is recommended that the authors explain the differences between the U-Net network used in their study and the classic U-Net, as well as the aspects in which their design differs.

Response: Thank you for this remark. The U-Net proposed by Ronneberger et al. was an asymmetrical network, i.e. the size of output map was different from the size of input image, whereas in our U-Net the size of input image was equal to the size of output map. The original U-Net performs binary segmentation, see Figure 3C in[36], whereas our U-Net performs multi-class segmentation. The next difference is that we employ Dice score/loss.

In the revised manuscript we added the following content:
In contrast to the U-Net proposed by Ronneberger et al. our network is symmetric one, i.e. the input image size is equal to output map size, it performs multi-class segmentation, and relies on the Dice loss and score for training and evaluation, respectively.

Comment #6: In the section of Radiographs Pre-processing, it is recommended that the authors provide an explanation of why they chose to use SDI to evaluate the accuracy of bone segmentation, instead of using mainstream segmentation evaluation metrics.

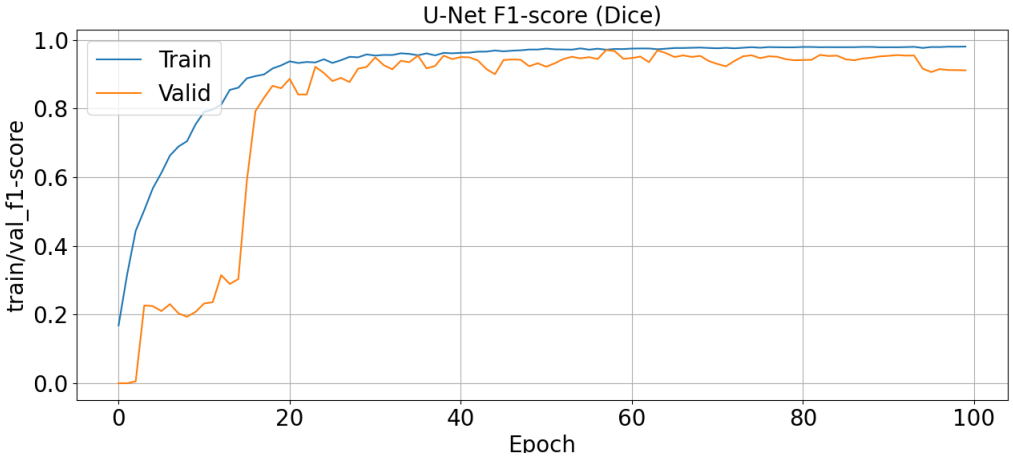
Response: Thank you very much for this comment.
We added the following explanation:
The accuracy of the bone segmentation was evaluated using SDI which is the most used metric in medical image segmentation[37, 39].

39 **Müller D**, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. BMC Research Notes 2022; **15**(1): 210[DOI: 10.1186/s13104-022-06096-y]

Comment #7: In the section of Architecture and training U-Net, the authors trained the model using 181 images and 60 epochs. Would this lead to insufficient training? It is recommended that the authors provide further explanation.

Response: Thank you very much for this valuable comment. We agree with you that a clarification in this point is needed. The hyperparameters of the final U-Net model were determined by trial and error. We trained several networks on our training/validation dataset to determine the parameters of the U-Net network, select the optimizer and determine the number of epochs. Almost all networks trained successfully at max. 60 epochs. For experimental purposes, we trained several

networks. In one case, the best Dice score was obtained at an epoch greater than 60, see the plot from this experiment below.



We modified the content of the manuscript as follows:

Original text:

It was trained in 60 epochs using Adam optimizer with Dice loss. Training consists of data augmentation using mirroring, rotations, and contrast enhancement.

Revised text:

The validation SDI was calculated at the end of each epoch during the training of the U-Net, and the training was stopped when the SDI did not increase over 10 following epochs. This served as an early stop technique to avoid overfitting, where the value of early stop (patience) was set to 10. The U-Net was trained using Adam optimizer with Dice loss, learning rate (LR) set to 0.0001 (with reducing LR on plateau) and batch size equal to 8. The number of epochs was set to 80, and a callback was used to save the best U-Net model and its weights. The training data was augmented using mirroring, rotations, and contrast enhancement.

Comment #8: In the experimental section, it is recommended that the authors present the experimental results and data in the form of tables.

Response: Thank you very much for this valuable comment. Below are the experimental results and data in the form of tables.

Table 2 Correlation of hallux valgus angle, intermetatarsal angle, and pre-operative surgical decisions between clinicians, and against AI (our algorithm)

	hallux valgus angle correlation (ICC)	intermetatarsal angle correlation (ICC)	pre-operative surgical decisions correlation
R-O_B	0.96	0.79	0.73 (61/84)
R-O_{A1}	0.96	0.81	0.62 (52/84)
R-OA_2	0.96	0.78	0.73 (61/84)
O_B-O_{A1}	0.96	0.91	0.75 (63/84)
O_B-O_{A2}	0.99	0.95	0.88 (74/84)
O_{A1}-O_{A2}	0.98	0.91	0.82 (69/84)
AI-O_{A2}	0.97 (AA-ICC) 0.97 (C-ICC)	0.89 (AA-ICC) 0.75 (C-ICC)	0.80 (67/84)

AI: Artificial intelligence; AA-ICC: absolute agreement interclass correlation coefficient; C-ICC: consistency interclass correlation coefficient; ICC: interclass correlation coefficient; O_{A1}, O_{A2}, O_B: orthopedic surgeons; R: musculoskeletal radiologist.

Comment #9: In the experimental section, it is recommended that the authors include experimental comparisons between the segmentation performance of the selected segmentation network used in their study and that of state-of-the-art segmentation networks.

Response: Thank you very much for this valuable comment. To the best of our knowledge, little work has been done in the area of bone segmentation of radiographs, excluding, to some extent, the task of chest segmentation. The problem of the foot bone segmentation differs from image segmentation, where a considerable work has been done. We experimented with several models, including Linknet, PSPNet, FPN, U-Net++, and U-Net Transformer. However, given the limited size of training dataset the results were worse. We also trained several U-Nets with encoders based on pre-trained backbones (on the imageNet dataset). U-Nets using encoders built on efficientnetbx, including efficientnetb7, and inceptionv3 achieved quite promising results. On the other hand, VGG networks obtained worse results. Given that our network is capable of achieving the Dice score > 0.97 the room for

improvement is quite small. Our initial results demonstrate, that the efficientnetb and inceptionv3 networks can be trained in smaller number of epochs, but the training time of single epoch is several times larger than time for training our network. Moreover, initial results demonstrate that results achieved by a U-Net built on inceptionv3 with random weights are only slightly smaller than results achieved by U-Net built on inceptionv3 pretrained on the imageNet dataset. This might suggest a limited usefulness of low-level features extracted by pretrained networks. There is no doubt that further research is needed. We tried to summarize this research but regarding the editorial constraints of WJO for number of words it was impossible to present and discuss such results. We are planning to continue research on this area. We will compare the performance and consider the explainability (also referred to as interpretability) of our networks vs. more complicated models, including calibration, which is itself a very important problem. Once again, thank you very much for your recommendation, which motivated us to compare the results obtained by our network with the results obtained by U-Net built on inceptionv3/efficientnet.