Comparing Segmentation, Classification, and Cascade Convolutional Neural Networks for Pneumothorax Prediction

Abstract

The diagnosis of many diseases, including pulmonary edema and pneumothorax, is heavily reliant on chest radiographs. Due to the restriction on the number of radiologists who are able to give diagnoses based on chest radiographs, disease treatment could be severely delayed. To solve this pressing problem, machine learning models like convolutional neural networks (CNNs) have been widely used to help give patients diagnoses based on their chest X-rays.

Last quarter, we focused on training CNN models, including VGG 16, 19, and ResNet 152, to give predictions for pulmonary edema based on chest X-rays. This quarter, we are interested in exploring more about using segmentation models to predict pneumothorax and comparing the performances among different model structures, including segmentation models, classification models, and cascades of segmentation model and classification model. We used a chest radiograph dataset called CANDID-PTX, curated by researchers from New Zealand, and annotated by senior consultant radiologists for ground truth labels. As expected, the cascade models perform the best among three structures, the segmentation models being the second, and the classification models being the least ideal.

Introduction

Deep learning algorithms, especially convolutional neural networks (CNNs), have been widely used in medical image interpretation. In recent years, multiple research groups have shown the applicability of CNNs in pulmonary disease detection via chest radiograph interpretation and classification[1-3]. One of the pulmonary diseases that shows promising potential for CNNs is pneumothorax.

We decided to use the CANDID-PTX dataset, with 19,237 anonymized patient chest radiographs from New Zealand[4]. With the manually annotated masks for each positive pneumothorax X-ray and a large amount of normal chest X-ray, we trained two classification models: ResNet 34 and EfficientNet-B3, two segmentation models with ResNet 34 and EfficientNet-B3 as encoder and UNet as decoder, as well as cascade models that pipe classification models after segmentation models. We are interested in investigating whether the cascade models will perform better than either one of the

components by itself, given that it should theoretically integrate the benefits of both segmentation and classification models.

Methods

Acquisition of the CANDID-PTX dataset

As required by the researchers who curated the CANDID-PTX dataset, we completed an online ethics course and signed the data usage agreement before obtaining the dataset.

Creation of training, validation, and test sets

The original dataset has 19,640 entries, with 19,237 unique chest radiographs. The difference is due to the fact that one radiograph might have more than one pneumothorax, making it more than one entry. During preprocessing, we selected all the unique chest radiographs and summed the masks with multiple pneumothorax if there are more than one. Among the 19,237 unique chest radiographs, there are only 3,196 positive cases, the rest being negative. We split the dataset into training set with 2,556 positive cases and 2,556 negative cases, validation set with 320 positive cases and 1,600 negative cases, and test set with 320 positive cases and 1,600 negative cases. We purposely made the training set balanced with positive and negative cases for better model training. For the validation and test set, we kept the 1:5 ratio of positive and negative cases to mimic the real-life situation where there are more negative cases than positive ones. In order to utilize all the negative cases, we replaced the negative cases in the training set with unused negative cases every four epochs. Due to GPU memory constraints, we resized all the images from resolution of 1,024 x 1,024 to 512 x 512.

Classification, Segmentation, and Cascade Model Training

Two classification models (ResNet 34, EfficientNet-B3), two segmentation models (ResNet 34/UNet, EfficientNet-B3/UNet), and four cascade models with classification models piped after segmentation models (ResNet 34/UNet + ResNet 34, ResNet 34/UNet + EfficientNet-B3, EfficientNet-B3/UNet + ResNet 34, EfficientNet-B3/UNet + EfficientNet-B3) were trained with the same training, validation, and test pipeline as mentioned in the previous section for 20 epochs. All models were initialized with ImageNet weights. During training, all segmentation models have all of their parameters trainable. For EfficientNet-B3, the first 6 layers were frozen for faster training and GPU constraints. For ResNet 34, we unfroze the last two layers and left the rest frozen for better recall. The loss function we used was binary cross entropy loss. The optimizer we used was Adam optimizer. The batch size was 4, and the learning rate was 0.0001. The threshold for each pixel to be classified as 1 was 0.3 for the segmentation models. The minimum activation size for the radiograph

to be classified as positive was 375 for the segmentation models. We referred to the results published by a separate study using the same dataset and similar segmentation models for the cutoff values, but scaled down to fit our smaller number of pixels per image[5].

After training each segmentation mode for 20 epochs with the CANDID-PTX dataset, we saved the weights and reloaded the model for the cascade pipeline for better comparison. For the cascade pipeline, we saved the predicted masks from the segmentation models, along with two channels of the original input chest radiographs. The three-channel images were then input into the classification model for pneumothorax classification.

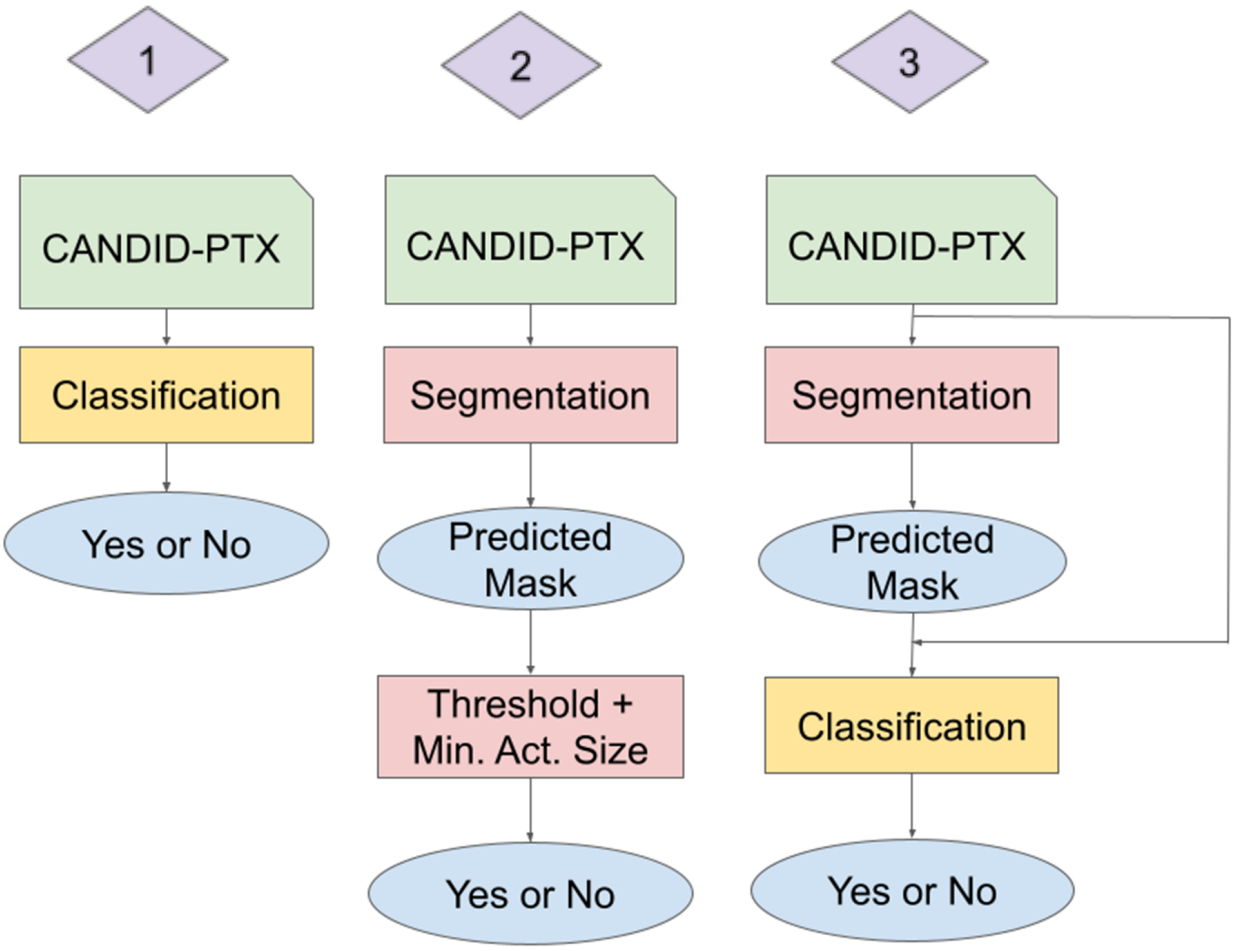

Figure 1: Illustration of different pneumothorax classification structures. The first model type is purely classification. The second model type is segmentation with hard thresholds to turn predicted masks into binary classification. The third model type is the cascade models with both segmentation and classification portions to avoid hard thresholds while still returning binary classifications as well as predicted masks.

Model Metrics

For the final model evaluation, we used F1 and recall scores, due to the fact that for medical images, it is less harmful to have false positive cases than false negative cases since radiologists will usually double-check with each positive radiograph before giving a confirmed diagnosis. Therefore, the best model should ideally have the highest F1 and recall scores.

Results

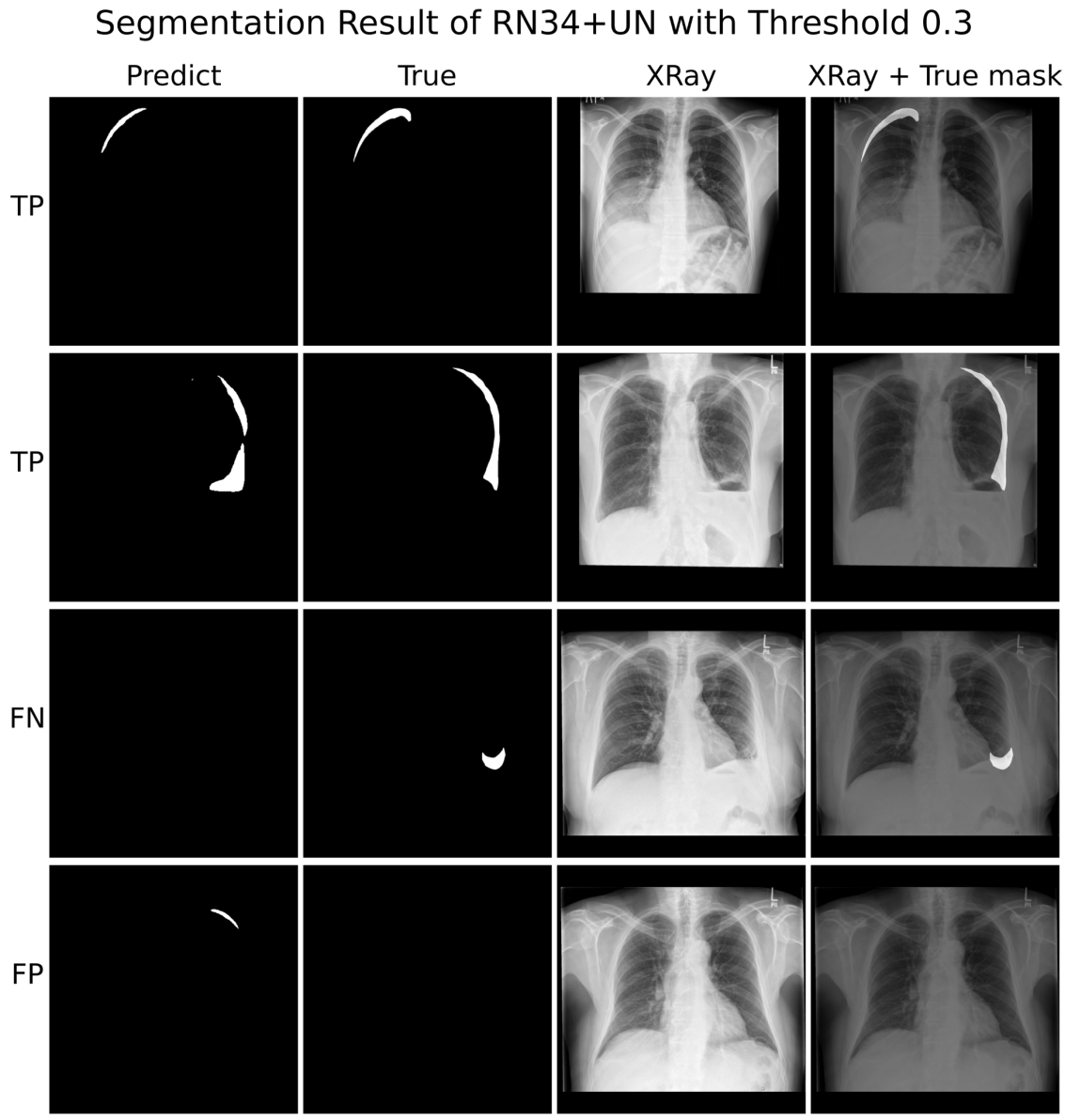

Figure 2: Sample Predicted Mask, True Mask, Chest XRay, and Overlaid Chest XRays from ResNet 34/UNet segmentation model. Sample predicted masks versus true masks based on trained ResNet

34/UNet segmentation model. The threshold for pixels to be classified as 1 was 0.3. TP: true positive, FN: false negative, FP: false positive.

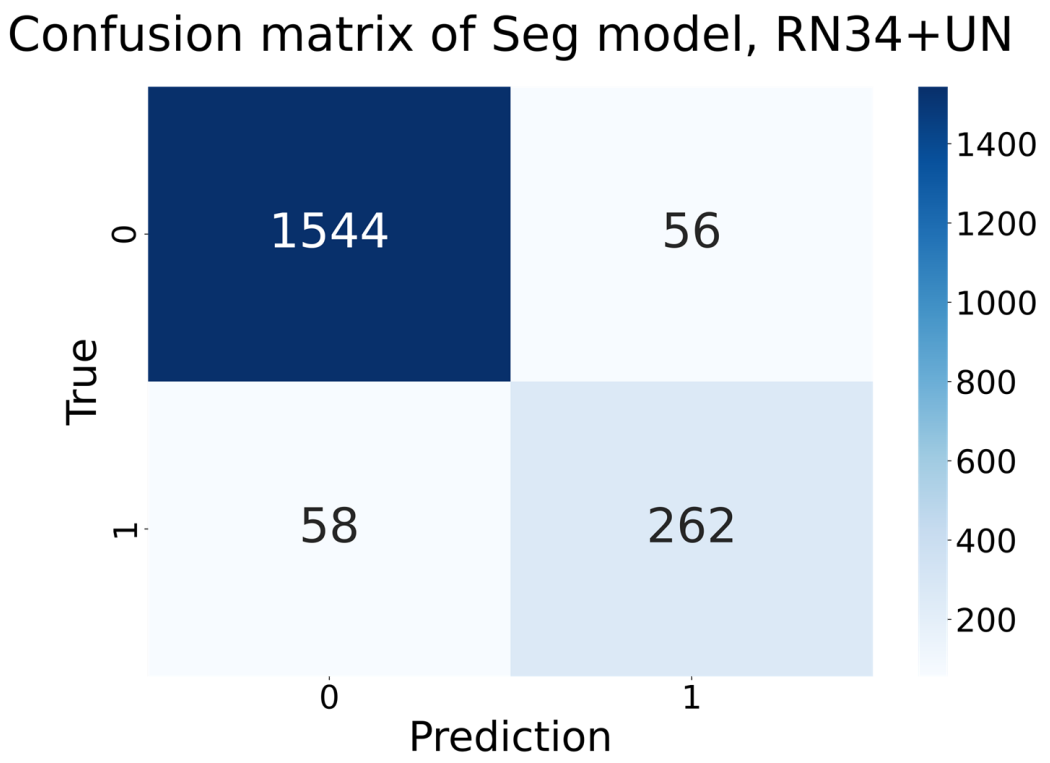

Figure 3: Confusion matrix for ResNet 34/UNet segmentation model. Based on the confusion matrix, the trained ResNet 34/UNet segmentation model classified 1806 radiographs correctly. The model made 58 false negative predictions and 56 false positive predictions.

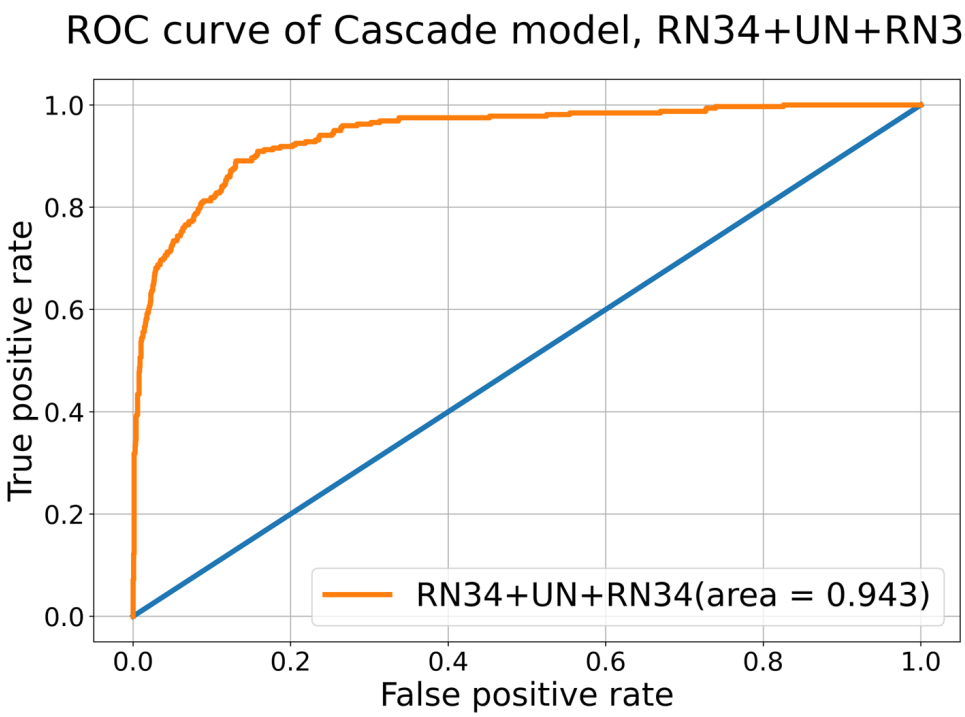

Figure 4: ROC curve for the cascade model: ResNet 34/UNet segmentation + ResNet 34 classification model. Demonstration of the ROC curve for one of the cascade models, with the ROC-AUC score being 0.943.

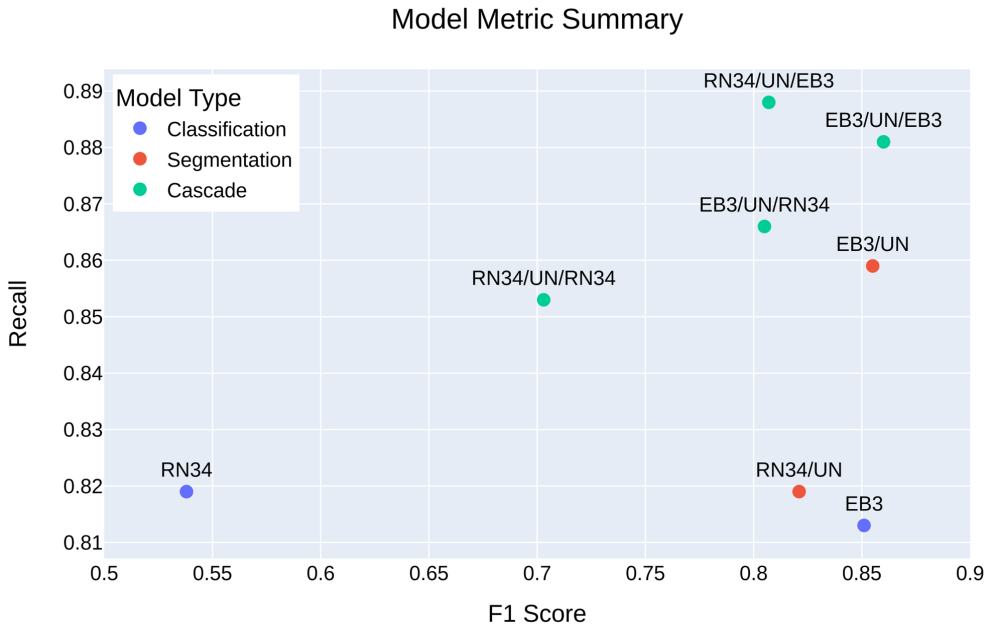

Figure 5: Recall vs F1 scores of all tested models. RN34: ResNet 34, EB3: EfficientNet-B3, UN: UNet, ResNet34/UN: ResNet 34 as encoder & UNet as decoder, EB3/UN: EfficientNet-B3 as encoder &

UNet as decoder, RN34/UN/EB3: cascade of ResNet34/UN segmentation model + EfficientNet-B3 classification model. All the structures with three labels have the first model as the encoder, the second model as the decoder, and the third model being the classification model in the cascade model.

Based on the metric F1 and recall score, the best model overall is the cascade model EfficientNet-B3/UNet segmentation and EfficientNet-B3 classification model. Overall, cascade models perform better than segmentation models, which outperformed classification models. Within the same structure, EfficientNet-B3 outperformed the ResNet 34 models.

In addition, we tested different strategies for dealing with the imbalanced dataset. Our first strategy was to train each model with all the positive and negative cases not included in the validation and test set, but the drawback was that the training set has too many negative cases for the model to learn how to find pneumothorax in the positive cases well. In addition, the training time was very long due to the large training set size. Our second strategy was to initially train the models with balanced positive and negative cases, and then replace the negative cases with unseen negative cases every several epochs. This turned out to be the best strategy both in terms of performance and training time. The third strategy was to initially train with balanced positive and negative cases, and then gradually add more negative cases to the training set. This strategy did not outperform the second strategy on the same model, and the training time was much longer.

Conclusion and Discussion

As expected, the cascade models had the best overall performance among the three different model structures. Specifically, segmentation models piped with EfficientNet-B3 performed the best. EfficientNet-B3 was a much better model than ResNet 34, since the segmentation model with EfficientNet-B3 as encoder outperformed cascade models with ResNet 34 as encoder piped with a separate ResNet 34 classification model. Due to rescaling of images from resolution 1024 x 1024 to 512 x 512, the true masks were no longer binary. In order to not introduce more biases, we did not plot the ROC curve for segmentation models. For future reference, if the true masks and training set are both in 1024 x 1024 resolution, ROC curves could be plotted for segmentation models, and there could be one more metric for all models. We calculated dice coefficients for the segmentation models, but since they do not apply to classification models, they were excluded from the final comparisons.

For all the models, we trained them for 20 epochs based on experience, and made sure that they were not overfitting too much from the validation loss. For cascade models with EfficientNet-B3 as the classification component, we did see a trend of overfitting a little, but the test metric results were not dramatically different when we trained with fewer epochs.

We concluded that cascade models that pipe segmentation and classification models together would be a good potential model for future pneumothorax classification. It has high F1 and recall score, and it generates both the predicted label and predicted pneumothorax masks without the trouble of finding the best thresholds and minimum activation size for segmentation models. We did not get the chance to generalize the model to other diseases, but it would be an interesting potential future project to test how robust the models are for various different diseases.

Citations

[1] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Auto- mated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, Aug. 2017, doi: 10.1148/radiol.2017162326.

[2] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Sci. Rep.*, vol. 10, no. 1, pp.1–12, Dec. 2020, doi: 10.1038/s41598-020-76550-z.

[3] E. J. Hwang, J. G. Nam, W. H. Lim, S. J. Park, Y. S. Jeong, J. H. Kang, E. K. Hong, T. M. Kim, J. M. Goo, S. Park, K. H. Kim, and C. M. Park, "Deep learning for chest radiograph diagnosis in the emergency department," *Radiology*, vol. 293, no. 3, pp. 573–580, Dec. 2019, doi: 10.1148/radiol.2019191225.

[4] Feng, Sijing et al. "Curation of the CANDID-PTX Dataset with Free-Text Reports." *Radiology. Artificial intelligence* vol. 3,6 e210136. 13 Oct. 2021, doi:10.1148/ryai.2021210136

[5] Feng, Sijing et al. "Automated pneumothorax triaging in chest X-rays in the New Zealand population using deep-learning algorithms." *Journal of medical imaging and radiation oncology* vol. 66,8 (2022): 1035-1043. doi:10.1111/1754-9485.13393