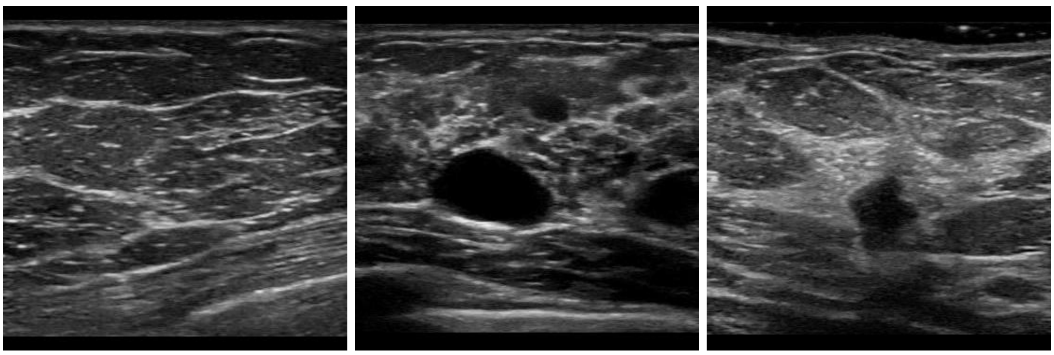
Uncertainty Quantification in Deep Learning for Breast Cancer Classification in Point-of-Care Ultrasound Imaging

Popular Science Summary

Marisa Wodrich

Did you know that your chances of survival of breast cancer are way lower in low- and middle-income countries? In Kyadondo, Uganda, barely 12% of breast cancer patients are still alive five years after diagnosis, whereas in many high-income countries more than 85% of people are still alive at that point. The medical tools and number of trained medical doctors we have in high-income countries like Sweden are not available in those countries. Solutions to improve the survival rates like mammography screenings become infeasible due to high costs in equipment and education. We therefore propose to use an AI-based approach to solve this problem. Images are taken with a small and cheap pocket ultrasound device and are then evaluated by an AI-based algorithm. The algorithm looks at the image and predicts if it is either breast cancer, a non-cancerous lesion, or just normal breast tissue. While this works quite well, the algorithm will still make some mistakes and predict a wrong class sometimes. Consequently, we do not want to blindly trust the prediction of the algorithm, especially not in cases where it would say that the patient is healthy, when in fact they have breast cancer and fast treatment would be needed for survival. Mistakes from our algorithm can have severe consequences, up to life-threatening outcomes, and therefore need to be minimized.



Examples of pocket ultrasound images. From left to right: normal breast tissue, benign lesion (non-cancerous), malignant lesion (breast cancer).

In this thesis, we looked into different methods that can be used to tell how certain or trustworthy a prediction from our algorithm is. This means that additionally to telling us what diagnosis it thinks an image should get, the algorithm also tells us how certain it is that it made the right choice.

We found that for all methods we used, having these uncertainty values alongside the predicted diagnosis helps the overall goal: Predictions were very likely to be true when the uncertainty value was low, and more likely to be wrong for higher uncertainties. This means that the uncertainty values could be used to determine how trustworthy a prediction was. By setting a threshold for the uncertainties, we were able to filter out many wrong predictions, with the remaining ones being mostly correct predictions. This therefore can be used in practice to increase the safety of our algorithm and make it more usable in the real world. The cases in which the uncertainty is too high would need further images being taken, or human intervention to determine the diagnosis.

Reasons for being uncertain can for example be that the algorithm has not seen enough similar images before (during its training phase) to confidently make a prediction, or that the image is of bad quality. It can also be due to the algorithm just not being the best fit for the actual task. Overall, there are many sources from which uncertainties can originate.

With most of the investigated methods for measuring uncertainties, we were also able to detect images that were very different from the usual images (and therefore often resulted in high uncertainties in the prediction). For example, we used images of handwritten digits and found that with the uncertainty values, we can filter images like that out quite well with most methods. This can especially be useful in practice when there are images that are very different to all images our algorithm has seen before. In those cases, it does not know how to interpret the image and would rather make some sort of random decision that we also do not want to trust. It is therefore extremely useful if our uncertainty methods can detect such images as well.

Overall, we tested three different AI-based methods, which mostly differ in how complex they are. For each of them, we tested different ways to calculate uncertainties. In general, the more complex and complicated methods perform better, but all methods work and could be used together with thresholds to determine the trustworthiness of a prediction.

With our good results we come to the conclusion that our methods show great potential to be used in a real-world medical setting. We suggest that the AI-based algorithm will make a first assessment, predicting a diagnosis, and one of our uncertainty methods is used to calculate how trustworthy the prediction is. In cases of high uncertainties, a medical professional should instead make the diagnosis and the algorithm should not be trusted, or new images should be taken. In other cases where the uncertainties are low, the diagnosis from the algorithm can most likely be trusted. These findings are a great baseline and prove the usefulness of taking uncertainties into account. Applying our project in countries like Uganda can potentially save many lives.