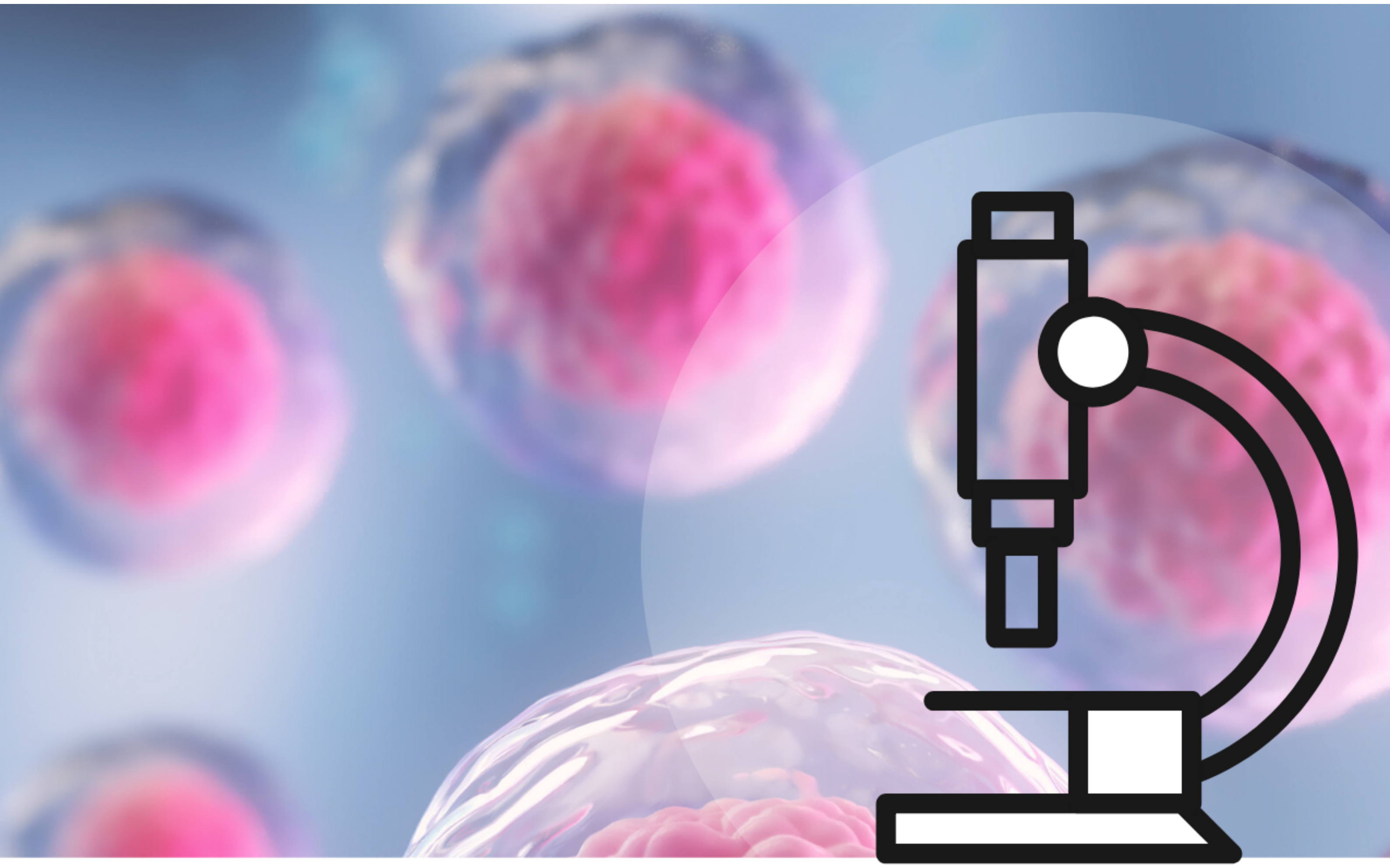


DEEPKAPHA.AI LABS PRESENTS

CYTODEEP FEASIBILITY STUDY

ASSESSMENT OF ARTIFICIAL INTELLIGENCE IN BREAST CANCER DIAGNOSTICS

Feasibility Assessment: CytoDeep Project

Technical Feasibility

Background rationale

In 2018, it was estimated that more than 600,000 people lost their lives to breast cancer (Globocan 2018). Early diagnosis is a critical component in treating breast cancer, but unfortunately for those living in developing countries it is often diagnosed at later stages than in developed countries causing higher rates of mortality (Coleman et al, 2008 and Unger-Saldaña 2014). In developing countries, where physicians are already in short supply, the current breast cancer diagnostic procedure is difficult and time consuming.

There are many potential factors which could cause this severe discrepancy in the rates of survival between developing and developed countries, but perhaps one of the most significant is the stage in which the cancer is diagnosed. In low-and middle-income countries (LMIC), it is estimated that only 20-50% of breast cancer diagnoses are made while the cancer is in stage I or II, but in high-income countries, this rate is at 70% (Unger-Saldaña 2014). It cannot be understated the importance of early diagnosis as it is a significant predictor of survival (Anderson, 2003).

The manner in which breast cancer is typically diagnosed involves a combination of several different tests typically in the order of clinical breast examination, diagnostic imaging, and then a biopsy. During a clinical examination, a physician will perform a palpable examination to determine if a lesion can be felt and then may also perform a mammogram or ultrasound. If a concerning lesion is palpable or detected through diagnostic imaging, a biopsy will be performed on the identified lesion. This biopsy is often fine-needle aspiration cytology (FNAC) or core-needle biopsy (CNB). CNB is known for its high diagnostic power and can be used for tumor grading. However, despite having slightly lower diagnostic performance and decreasing usage in developed countries, FNAC is a much cheaper procedure making it ideal for application in developing countries.

In the meantime, artificial intelligence (AI) is showing promise to help remedy this issue of diagnosis as some machine learning (ML) and deep learning (DL) models have already assisted in clinical decision making or matched the diagnostic accuracy of trained physicians (Liu et al 2019). Nevertheless, these diagnostic models do have some significant limitations such as needing significant amounts of labelled training data and possessing a relative lack of explainability. To address these challenges, our proposal is to implement an AI model which will act as a breast cancer cytological diagnostic aid for physicians in the developing world. In detail, the technical, financial and legal aspects will be discussed to show how the model can be implemented in a manner which is fast, accurate, and relatively inexpensive.

Introduction

It is clear that breast cancer is one which affects many worldwide with more than 2 million cases diagnosed in 2018, making it the second most common form of cancer (Globocan 2018). Around the world, breast cancer is the second most common form of cancer and is the most frequent form of cancer in women by a significant amount (Ferlay et al 2015). The rates of incidence are the highest in developed regions at 74.1 per 100000 compared to 31.3 in developing regions (Ferlay et al 2015). The rates of mortality, however, are much higher in LMIC and as an example, the female rate of survival for breast cancer in Algeria (38.8%) was nearly half that of the rate of the average European country (73.1%) (Coleman et al 2008). In developing countries, where the number of trained pathologists is often scarce, it is reasonable to investigate how AI can act as a diagnostic aid to reduce the workload and increase the objectivity of diagnoses performed by pathologists.

Recently, AI has shown to be a very powerful tool in many various domains such as computer vision, natural language processing, and autonomous driving. Given sufficient data, AI attempts to mimic complex cognitive functions and possess the capability to learn from experience allowing its performance to improve over time. With data being of fundamental importance to the field of healthcare combined with the powerful ability of AI to extract meaningful representations from data, it is apparent that AI has the potential to bring about a paradigm shift in healthcare.

Some tasks which were previously believed to have been only possible by highly skilled physicians have now been accurately replicated by AI. In the field of deep-learning, which is a sub-field of AI, diagnostic models are already approaching the accuracy of health-care professionals and have been applied in medical domains such as cardiology, oncology, hepatology, and many others (Liu *et al* 2019). A meta-analysis by Liu *et al.*, 2019 compared the performance of 14 deep-learning models in diagnosing diseases with that of health-care professionals. They found that the models had a pooled sensitivity of 87.0% while the health-care professionals had sensitivity of 86.4% , and that the models had a pooled specificity of 92.5% while the health-care professionals had sensitivity of 90.5% (Liu *et al.*, 2019).

There are many areas of healthcare where AI has potential applications, but medical imaging is a field which is of great diagnostic importance to physicians. Medical image interpretation, depending on the domain, can be a time consuming and difficult process for physicians. With the potential speed and accuracy benefits of AI, there has been significant research on using deep-learning models for diagnosis from medical images (Lee 2017) and some models have already seen usage in the healthcare industry. The United States Food and Drug Administration (FDA) has already approved several medical imaging AI models to be deployed in a healthcare setting including diagnosing wrist fractures from x-ray, diabetic retinopathy, and strokes from computerized tomography (CT) scans (Topol 2019)

Before building a diagnostic model, it is important to understand the advantages and disadvantages of the procedure by which the model is basing its decision on. Breast cancer can be diagnosed by imaging such as an ultrasound and/or through a biopsy such as FNAC or CNB. In developed countries, the usage of FNAC for breast lesions has been reduced in favor of CNB for a variety of factors such as difficulty in diagnosis, but FNAC still has key advantages in its ease of use, inexpensive cost, and minimal infrastructure requirements (Hukkinen et al 2008).

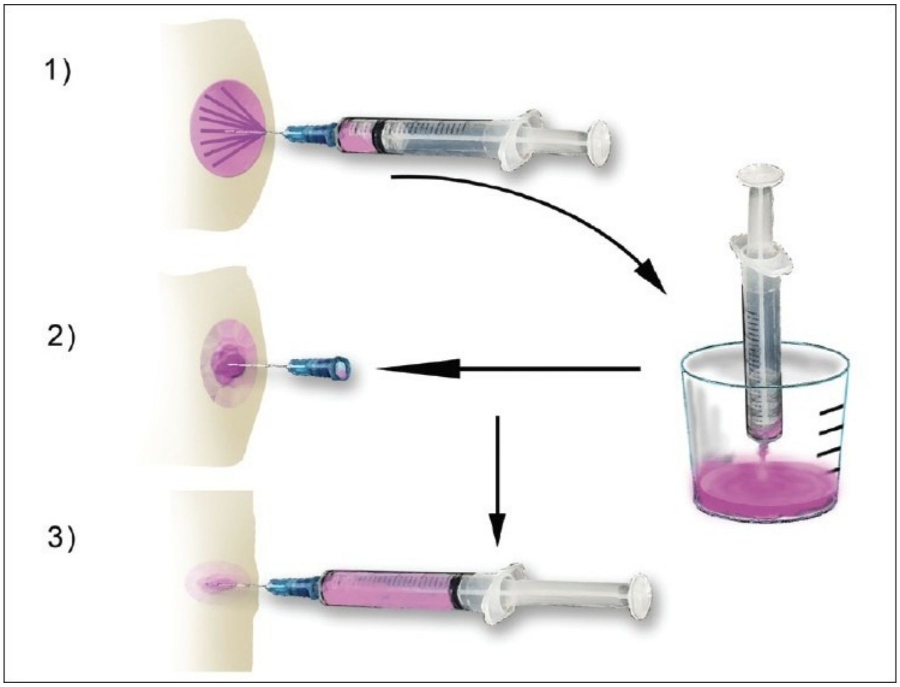
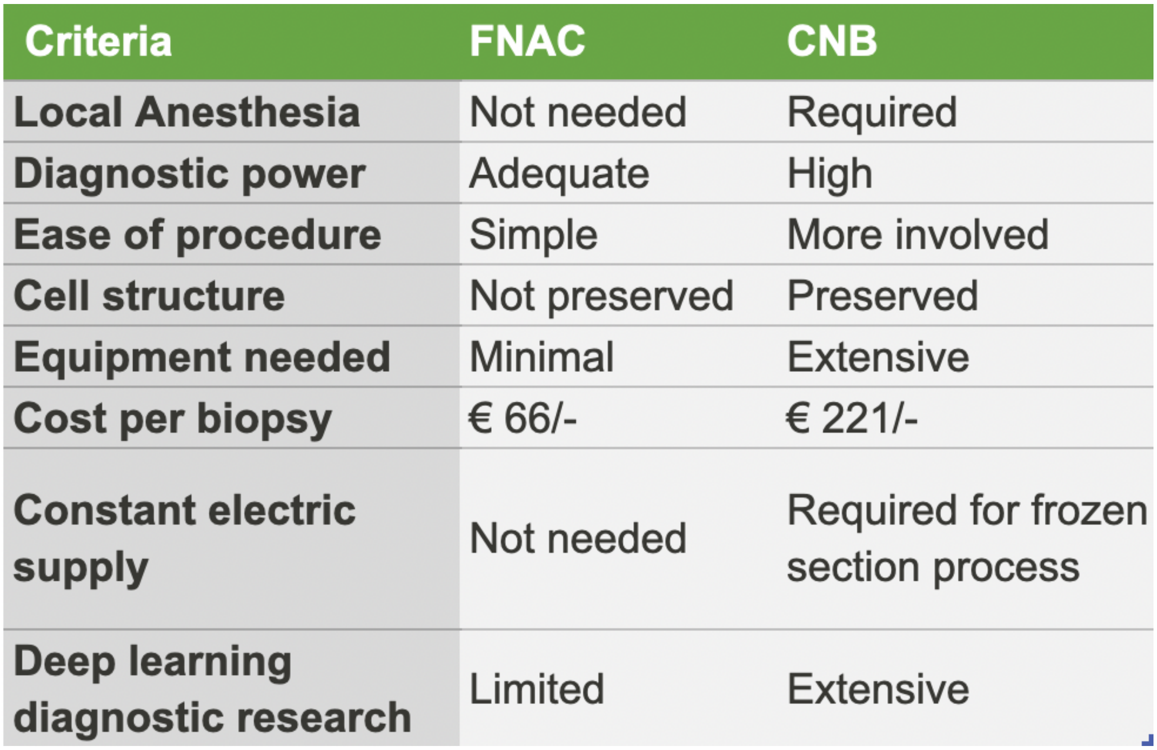


Figure 1. Diagram showing the process of performing fine-needle aspiration cytology on a lesion

One of the challenges of FNAC is the ability for a pathologist to make a definitive diagnosis. Accurate interpretation of FNAC results takes very extensive training and can still be somewhat inconclusive even for skilled physicians (Willems 2012). CNB results are generally easier for pathologists to diagnose and usually have better performance metrics. Although FNAC diagnosis does have very high sensitivity and specificity (Gary et al 2010), a meta-analysis conducted by Willems *et al.* showed that CNB has higher sensitivity and specificity than FNAC for identifying breast lesions. Samples which are extracted via CNB also preserve the cell structure of the lesion unlike FNAC. This allows for histological diagnosis which can give pathologists more insight into precisely which type of breast cancer the patient may have. Histopathology also allows for tumor grading, unlike cytopathology where there is not a definitive consensus of what criteria can be used to grade FNAC results (Khan 2003).

Despite the limitations of FNAC, it has a few key advantages over CNB which are especially pertinent to its usage in developing countries. It is a minimally invasive and straightforward procedure with low risk of potential complications. FNAC does not require anesthesia whereas CNB requires local anesthesia (Łukasiewicz et al 2017). Additionally, FNAC results can be prepared very quickly allowing for same day diagnosis (Willems et al 2017). This can help reduce the associated stress of the patient faced with a potential cancer diagnosis.

Perhaps the most significant disadvantage of CNB compared to FNAC is the large cost associated with CNB. Preparation of histological tissue requires more specialized equipment including microtomes to cut sample tissue (Ross et al 2010). Additionally, if frozen section processing is used it demands highly skilled histo-scientists prepare the tissue inside a subfreezing chamber called a cryostat, and one of the primary drawbacks to this method is that the operation of a cryostat requires a constant supply of electricity (Ross et al 2010). On average, the average cost of performing a biopsy with CNB (221/-) has been shown to be more than three-times that of FNAC (66/-) (Vimpeli 2008). For those in rural regions of developing countries the lack of electrical infrastructure, more expensive equipment, and additional skilled technicians needed makes implementing diagnosis of breast cancer with CNB infeasible.



Criteria	FNAC	CNB
Local Anesthesia	Not needed	Required
Diagnostic power	Adequate	High
Ease of procedure	Simple	More involved
Cell structure	Not preserved	Preserved
Equipment needed	Minimal	Extensive
Cost per biopsy	€ 66/-	€ 221/-
Constant electric supply	Not needed	Required for frozen section process
Deep learning diagnostic research	Limited	Extensive

Figure 2. Table comparing fine-needle aspiration cytology (FNAC) and core-needle biopsy (CNB) for breast cancer diagnosis. FNAC has key advantages in equipment required and lower cost over CNB.

The application of AI can also be used in breast cancer diagnosis where deep-learning models have been applied to mammograms, histology, and cytology. The table below shows the performance of different breast cancer diagnostic models, and for some of the studies, the test set was also evaluated by health-care professionals. This allows for a performance comparison between the deep-learning model and health-care professionals and many of the models have performance metrics approaching or even exceeding that of health-care professionals. The inclusion criteria included deep learning models used for breast cancer diagnosis or classification which had also had their performance compared with health-care professionals. Also, two relevant studies which used deep-learning models specifically for breast cancer diagnosis from FNAC results were included as well, however these two models were not compared against health-care professionals.

Reference	**Size of Dataset (Train/Test)**	**Data**	**Algorithm**	**Model Performance**	**Health-care Professional Performance**
Byra et al 2018	n = 1,032 (training: 85.5%)	Breast ultrasound images	VGG-19 (Transfer Learning)	AUC = 0.936	3 Radiologists had ranging AUC = 0.806-0.882
Becker et al 2018	n = 637 (training: 69.9%)	Breast ultrasound images	ViDi Suite Version 20	AUC = 0.84 Specificity = 80.4% Sensitivity = 84.2%	Radiologist had AUC = 0.89 Specificity = 89.4% Sensitivity = 84.2%
Bejnordi et al 2017	n = 399 (training: 67.7%)	Breast lymph node histology photographs	GoogleLeNet, ResNet, VGG-16, VGG-Net, U-net, and others	AUC = 0.994 (Mean AUC of the top 5 models = 0.966)	11 Pathologists under time constraint had mean AUC = 0.810 and without time constraint had mean AUC = 0.960
Fujioka et al 2019	n = 1,087 (training: 90.0%)	Breast ultrasound images	Inception V2 (Transfer Learning)	AUC = 0.913 Specificity = 92.5% Sensitivity = 95.8% Accuracy = 92.5%	3 Radiologists had ranging AUC = 0.728–0.845 Specificity = 60.4% –77.1% Sensitivity = 72.8% –84.5% Accuracy = 65.8% –79.2%
Zejmo et al 2017	n = 50 (training: 60.0%)	Breast cytology images (FNAC)	AlexNet, GoogLeNet	Accuracy = 83%	N/A
Garud et al 2017	n = 37 (training: 54%)	Breast cytology images (FNAC)	GoogLeNet	Accuracy = 89.7% PPV = 85.48% NPV = 92.96%	N/A

Table 1. Models for the diagnosis of breast cancer. AUC: Area under the curve; PPV: Positive predictive value; NPV: Negative predictive value; FNAC: Fine-needle aspiration cytology; N/A: not available.

The above table can help serve as a baseline for the expected performance of various deep learning algorithms which may be appropriate for the proposed model. Many of the highlighted models had performance metrics which were comparable to that of health-care professionals. As anticipated, the majority used a convolutional neural network (CNN) which has become the most common deep-learning architecture to be used with image classification.

Although training deep-learning models such as CNNs can be quite memory intensive, many of the above studies which reported their hardware usage trained their models on commonly available graphics processing units (GPUs). Both Becker *et al.,* 2018 and Fujioka *et al.*, 2019 used the Nvidia GeForce GTX 1080 while Zejmo *et al.*, 2017 used the Nvidia GeForce GTX TITAN X with 12GB of RAM.

There are several papers on deep-learning models which have been made for histopathological or ultrasound imaging breast cancer diagnosis, but there is a relative lack of research for breast cancer cytological diagnosis using deep learning models (Litjens *et al.,* 2016 and Araújo *et al.,* 2017 and Cruz-Roa *et al., 2017*). To our knowledge as of February 2020 there have only been two published papers which have applied deep learning models to breast cancer cytology diagnosis (Zejmo *et al.*, 2017 and Garud *et al.*, 2017). Despite having relatively small datasets, the models by Zejmo et al 2017 and Garud et al 2017 had accuracy of 83% and 89.7% respectively which is comparable to the estimated 88.0% accuracy for pathologists (Yamaguchi *et al.*, 2012).

With the papers by Zejmo *et al.*, 2017 and Garud *et al.*, 2017 being the primary works available on the subject of deep-learning models applied to FNAC breast cancer diagnosis, these models will serve as a baseline for expected model performance. We have a host of potential improvements that can be employed to increase their performance which will be discussed at length in the Technical Feasibility section.

Technical feasibility

Participants and data

For this study, we have been working with several parties – both medical consulting organizations as well as data platform providers that consist of rich medical data in the form a web application service. Some platform providers offer datasets online while others have proprietary data and offer their services as telemedicine and telemonitoring services.

Dataset, assessment and exploratory data analysis

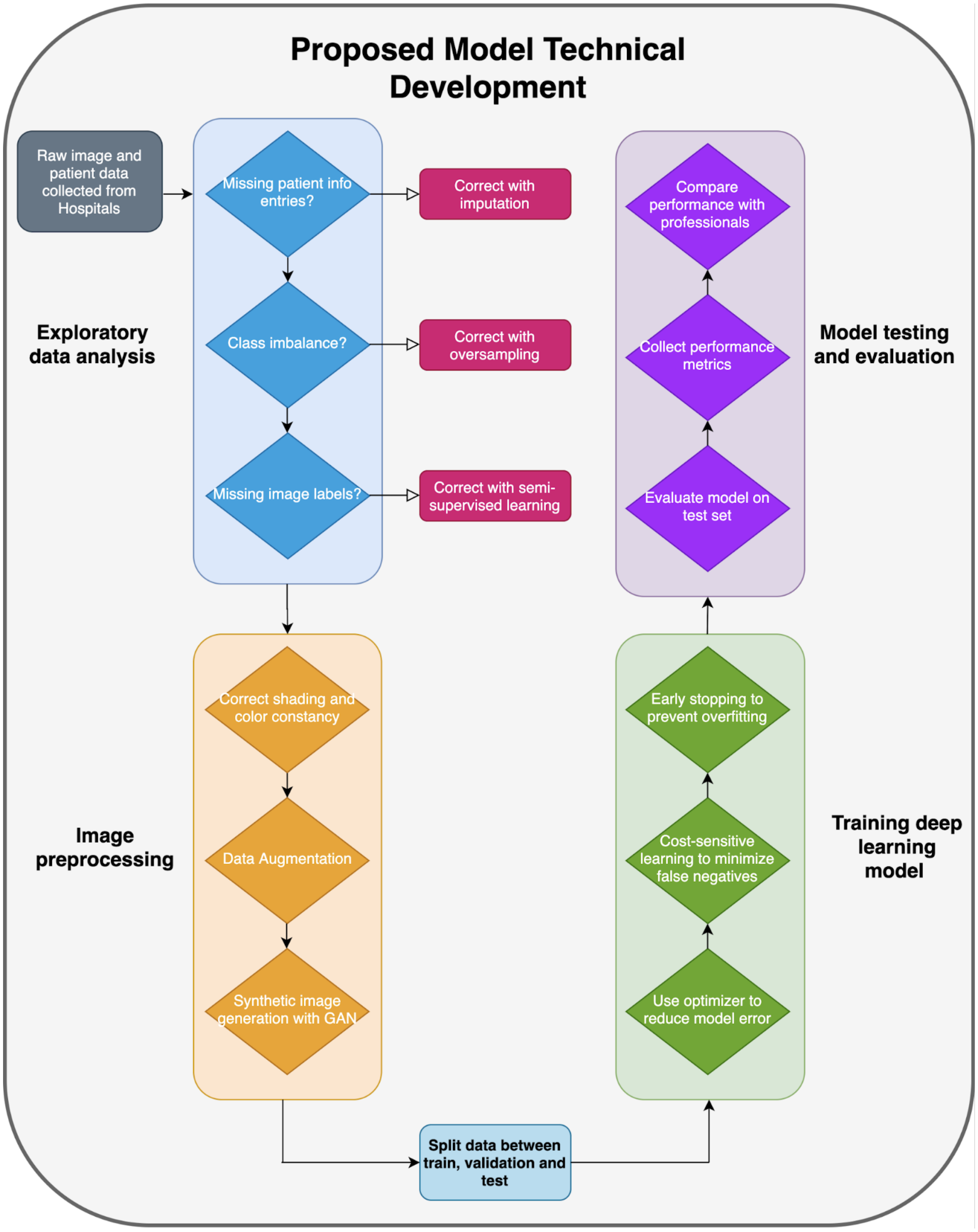

Proposed Model Technical Development
Raw image and patient data collected from Hospitals
Exploratory data analysis
Missing patient info entries?
Correct with imputation
Class imbalance?
Correct with oversampling
Missing image labels?
Correct with semi-supervised learning
Image preprocessing
Correct shading and color constancy
Data Augmentation
Synthetic image generation with GAN
Split data between train, validation and test
Training deep learning model
Use optimizer to reduce model error
Cost-sensitive learning to minimize false negatives
Early stopping to prevent overfitting
Model testing and evaluation
Evaluate model on test set
Collect performance metrics
Compare performance with professionals

Figure 3. This is the flowchart of how the proposed model will be developed. First we will perform exploratory data analysis on the dataset to determine which techniques will be needed to ensure the model will perform well. Then, we will preprocess the images and split them between the train, validation and test sets. Lastly, the model will be trained and its performance will be compared to that of healthcare professionals.

Exploratory data analysis

Prior to training, performing some exploratory analysis on the dataset can give some important insight into the amount of missing data and the distribution of classes to which the data belongs. This analysis is necessary in order to determine which techniques can be employed in order to correct for issues in the dataset.

Encountering missing data is typical when dealing with medical datasets as if it is being collected from different sources, not all records may have the same features. The information which accompanies the images such as patient sex, age, and medical history may have missing values, however, the cause for why they are missing can affect which methods can be applied to correct for this. There are four primary categories of missing data: missing completely at random, missing at random, missing that depends on unobserved predictors, and missing that depends on the missing value itself (Gelman et al 2006). If the data is missing completely at random, entries have equal probability of being missing, and in this scenario, removing missing values does not result in any bias being introduced into the model. If the data is missing at random, then the entries which are missing can be explained by the other attributes, so they can also be removed so long as variables that affect the probability of missingness are accounted for in the model (Gelman *et al.*, 2006). If the data is missing such that it depends on unobserved predictors, the probability that a given unit is missing can not be explained by the other attributes, so removing these entries will result in some bias being introduced into the model (Gelman *et al.*, 2006). Missingness that depends on the value of the missing value itself also will introduce some bias into the model if these entries are removed (Gelman *et al.*, 2006).

Since removing data entries which are not missing completely at random or missing at random can result in some bias being introduced into the model and decreases the size of the dataset, it may be preferable to use some form of imputation to predict missing entries (Gelman *et al.*, 2006). Imputation involves using some form of estimation based on other entries in the dataset to estimate missing values and there are varying levels of imputation complexity (Gelman et al 2006). A simple procedure to replace a missing continuous entry would be to replace it with the mean value for that attribute, but this has the drawback of decreasing the variance of the attribute (Gelman et al 2006). A more complex model of imputation for when data is missing at random is to use some form of regression to estimate the missing entry (Gelman et al 2006). This can be done by taking the other attributes which are correlated with the missing value and using regression to predict the value of the missing entry. A variety of imputation strategies is discussed in detail in Gelman *et al* 2006.

Another important data analysis tool is the count plot which can show the distribution of the frequency of appearances in different classes. The seaborn python library has built in functionality to easily create count plots which can be used to evaluate class imbalance. Class imbalance pertaining to the proposed model may involve the majority of the examples in the dataset being breast FNAC images which are non-cancerous with only a relatively small portion of the dataset consisting of cancerous images. If this imbalance is not corrected for, it can lead to the model becoming biased towards predicting the majority class. To combat this, a variety of oversampling and undersampling techniques can be utilized and these methods are described in detail in the “Data Imbalance Challenges - Oversampling/Undersampling” section.

Image pre-processing

Before training a CNN on a given image dataset, it is typical to apply some pre-processing techniques to ensure consistency across all of the training examples. The structure of CNNs typically requires that the dimensionality of images remain consistent, and since medical images may come in various sizes and aspect ratios, they must be scaled to be of consistent size. This can be achieved by first analyzing the dataset to determine the most common image size and aspect ratio. Then, once the most common aspect ratio has been determined, the images can all be cropped to match this aspect ratio. After this, the width and height of images can thn be scaled so they are of consistent size.

Size of dataset ("small dataset" problem in medical science) and imbalanced data

The primary challenge with training a supervised deep learning model for medical image classification is the need for a large, well labelled, and balanced dataset. However, when working with real-world datasets, especially in the medical domain, it is often the case that they are small (e.g., in the diagnosis of breast cancer n<1000), imbalanced, poorly-labelled, having missing data, while obtaining more labelled training data is tedious and expensive. In the case of the proposed model, acquiring additional labelled training data is difficult as it requires a pathologist manually reviewing images of breast cytologic smears. Despite the limitations that often accompany real world medical datasets, there are still a variety of techniques which can be employed that allow machine learning models to be trained on less than ideal conditions. We have access to a labelled dataset which is sufficient in size that when combined with techniques such as regularization, data augmentation, oversampling/undersampling (Lemaître *et al.*, 2017), cost sensitive learning, imputation (Jerez 2010, Pigott 2001) , and transfer learning (Torrey 2010), could still provide improved diagnostic performance. The following section will describe each of these techniques in detail and how they can be applied to the proposed model to make it more robust and increase its general performance in the medical domain.

Prevention of learning noise - regularization

When training a deep neural network with many layers on a small medical image dataset, an area of concern is overfitting. This is when the model begins to learn the minor and irrelevant characteristics of the images instead of learning the dominant statistical patterns. In the case of an overfit model, it has learned details which are not relevant to the actual classification of the images which causes it to have poor performance on data different from that which it was trained on.

To combat this issue of overfitting, it is important to employ some form of regularization. Regularization can broadly be described as a group of techniques used to prevent machine learning models from overfitting. This can prevent the model from learning the noise of the images in the dataset and instead pick up on the general structure of the data. A common regularization technique applied to neural-networks is dropout. This is where some nodes in a layer of the neural-network are randomly shut off in a given training step but can be reactivated in the subsequent training step. It effectively distributes the classification decision of the network across more nodes as some portion will be shut off for a given training set. This prevents the neural network from relying heavily on the output of a select few nodes and makes it more robust. Although this is a relatively simple technique, it has been shown to significantly increase the general performance of neural networks on a variety of tasks (Hinton 2012). There are also other common regularization techniques such as L1 and L2 regularization which can prevent the neuron weights from getting too large.

Artificial expansion of dataset - data augmentation

Although regularization techniques like dropout are very useful, for tasks with small datasets paired with deep neural networks, this alone does not satisfactorily deal with the issue of overfitting. Since procuring additional training data would require pathologists to manually review images of cytologic smears collected from breast lesions, it is reasonable to conclude that increasing the size of the training set in this manner is infeasible. However, there is a manner in which the dataset can be artificially increased in size using data augmentation.

Data augmentation is a powerful regularization technique which allows the generation of new images based off from existing images in the training set, effectively allowing the dataset to increase in size and diversity. The process works by first taking an existing image in the training set and making some slight modification which does not change the class it belongs to. Then, this augmented image is trained on the model using the same label as the original image. Now the model has learned a different variation of the image allowing for the model to be more robust. This is a type of artificial expansion of the dataset where there are many possible combinations of augmentations which can be applied to an individual image thereby increasing the number of training instances greatly.

The specific methods used in data augmentation vary, but common image modifications include cropping, rotation, translation, scaling, brightness adjustment, and altering contrast or saturation. As an example, rotating an existing image in the dataset gives the model a new training instance of an object in a different orientation. If the model then encounters an image similar in orientation to this rotated image when it has been deployed in the healthcare setting, it will have a greater likelihood of making a correct prediction. Thus, data augmentation essentially increases the general performance of the model using nothing more than the existing training data (Perez *et al.,* 2017). In regards to the breast cytology images, a varying combination of the listed augmentation techniques can be applied to the images during the training process.

It is important to consider that there may be cases in which applying specific augmentation techniques may not be applicable to certain datasets, such as applying reflection to objects which are not symmetric. If it is known that a given augmentation may change the class which an image belongs to, this should not be performed. Additionally, the image augmentation process is typically applied to each individual image during a given training step. The augmented images are usually not saved in order to minimize memory and network demands during training.

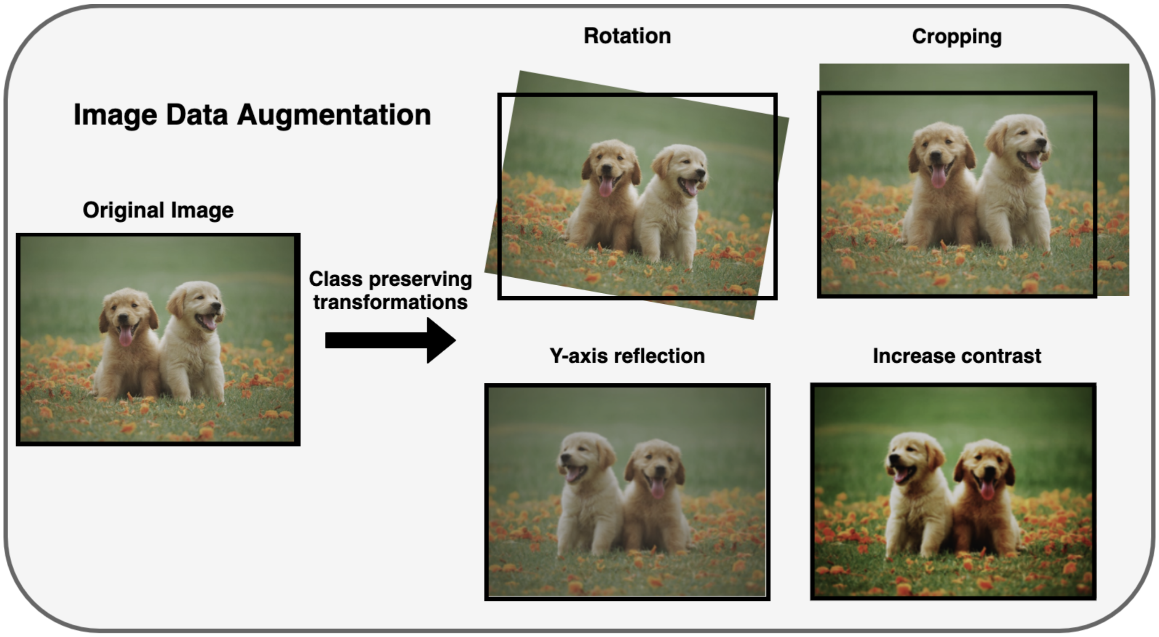


Figure 4. An example of possible transformations which can be applied to images during the data augmentation procedure.

Data imbalance: oversampling/undersampling

When working with a medical image dataset used for diagnosis, it is often the case that the dataset will be imbalanced, meaning the proportions of images belonging to certain classes is unevenly distributed. In context of the proposed model, the dataset of breast cytology images may consist of a higher proportion of non-cancerous images as opposed to cancerous images. Directly training a model on an imbalanced dataset can lead to poor performance but implementing oversampling and undersampling techniques can help address this issue.

Before addressing the methods which can be used to combat training a model on an imbalanced dataset, it is useful to first understand what may occur if no method is employed to deal with imbalance. Take the somewhat extreme example of a breast cytology dataset in which 95% of the images are non-cancerous and the remaining 5% are cancerous. Since the dataset is strongly skewed towards the majority class, the model could simply learn to classify all images which it is presented with as non-cancerous. This hypothetical model would have achieved a very high accuracy of 95% yet would be of little diagnostic value for a pathologist. Thus, when evaluating model performance, it is important to look at other metrics such as sensitivity, specificity, positive and negative predictive value, etc. particularly when the dataset is imbalanced.

To deal with imbalance, a simple technique which can be used is to increase the relative amount of examples in the underrepresented class which the model is trained on. Take the previous example where 5% of the data are in the minority class of cancerous images. We could randomly sample a disproportionate amount (e.g., 30%) of the training examples from this minority class and the remaining amount would be in the majority class. This would help balance the ratio of cancerous to non-cancerous images which the model is trained upon.

This technique is known as random oversampling, but it has a significant limitation in that training examples are being duplicated from the minority class which can cause the model to overfit (Fernández *et al.,* 2018). Conversely, random undersampling is where observations are randomly removed from the majority class, but this has the obvious drawback of losing potentially valuable training examples (Fernández *et al.,* 2018). A variety of different techniques to combat class imbalance can be easily implemented using the "imbalanced-learn" Python library (Lemaitre *et al.*, 2017).

To prevent the duplication of training examples caused by random oversampling, there are possible methods which generate new images from the minority class. For the proposed model, simple data-augmentation techniques (eg: rotation, reflection, etc.) can be applied to the cancerous breast cytology images to make new training examples for the model. There has also been research in using a generative adversarial network (GAN) to make new images for a model to train on (A. Antoniou, 2017). A GAN could take examples from the minority class of cancerous breast cytology images and generate entirely new and realistic examples for the model to train on. This is an example of synthetic data augmentation, where instead of simply taking an existing image in the training set and modifying it, we are generating an entirely new image which is similar to those in the training set.

These images generated using a GAN can have a much broader set of augmentations than traditional data augmentation techniques, giving a more diverse set of training data for the model (Antoniou, 2017). As an example of its application, this method of synthetic data augmentation with a GAN has been shown to significantly improve model performance in the classification of liver-lesions from CT scans, increasing sensitivity from 78.6% to 85.7% and sensitivity from 88.4% to 92.4% (Frid-Adar, 2018).

Controlling model error - cost-sensitive learning

When making a medical diagnostic test, some amount of error is unfortunately inevitable, however not all diagnostic errors have equal ramification. In the case of the proposed model which classifies breast cytology images as cancerous or non-cancerous, the potential errors which the model can make fall into one of two classes. Either the image is non-cancerous but the model classifies it as cancerous which would be a false-positive, or the image is cancerous but the model classifies it as non-cancerous which would be a false-negative. In the case of a false-positive result, the physician could rule out the error of a cancerous diagnosis with subsequent medical tests. However, a false negative may result in the physician believing the patient does not have cancer and thus may not follow up with additional diagnostic tests. This could lead to a potentially fatal outcome for the patient, so it is of great importance to try and minimize the potential of false-negative errors when designing a medical diagnostic model (Ling *et al* 2008). Cost-sensitive learning is a possible method which can address this issue of taking into account the unequal risk associated with different diagnostic errors.

The way machine learning classification models are typically optimized is to simply minimize the misclassification error. In the context of the proposed model, this means that the model would simply attempt to minimize the total number of incorrect classifications. This is described as cost-insensitive learning where the penalty the model receives for an incorrect classification is independent of the error type. However, in a medical diagnostic test it is known that making a false-negative error is a much more costly mistake, so instead of the model treating the errors equally, it can be optimized so that the model is penalized more for a false-negative error. Cost-sensitive learning is where the costs associated with specific types of misclassification are weighed heavier than others and has been shown to be useful in the medical domain (Park *et al.,* 2018).

Cost-sensitive learning is not only useful for accounting for different misclassification costs, but also can help with the issue of imbalanced datasets (Chawla *et al.*, 2000). Take the previous example where the dataset is highly skewed towards the majority class of non-cancerous breast cytology images. If the model is trained on this imbalanced dataset with cost-insensitive learning, it will likely become biased towards simply predicting that a presented image is non-cancerous. However, if a cost-sensitive approach is implemented, then the model will be penalized more for incorrectly classifying the cancerous breast cytology images which are in the minority class. Combining cost-sensitive learning with oversampling has been shown to be an effective way for dealing with the issue of class imbalance (Nguyen, 2010).

Addressing data labeling - semi-supervised learning

When working with medical image datasets, it is often the case that some samples may not have been adequately labelled by a physician. In the case of cancer diagnosis from biopsy, it is typical for a physician to take several samples from one patient and then review them under a microscope to determine if any are cancerous. Once the pathologist who is examining the tissue samples encounters a single slide they diagnose as cancerous, they may stop examining the remaining samples leaving some unlabeled. However, if the patient does not have cancer, it will require the physician to examine all of the tissue samples to affirm that all are non-malignant. A semi-supervised learning approach can prove to be beneficial in the case in which significant portions of the dataset are not adequately labelled.

A commonly used form of semi-supervised learning is self-training. Self-training is where the model is first trained on the portion of the dataset which is labelled well. Then after the model has achieved adequate performance on the labelled portion of the dataset, it is then used to predict the associated labels of the images which are missing labels. If the model has a high degree of confidence that the unlabelled image is being correctly classified, then this image is added to the training set with the associated label predicted by the model along with the existing data that is properly labelled. This process can be repeated until all of the unlabelled images have been categorized by the model.

Machine Learning

Recently, the most successful AI models have been machine learning based with their primary advantage being the associations and patterns which they detect do not have to be explicitly programmed. In years past, so called “expert systems” required domain experts to write pre-programmed instructions for AI to make decisions upon, but these systems had many notable limitations. Expert systems demanded domain experts write complicated sets of rules for the model to base its decisions on and these systems could not improve their performance over time without external input. Additionally, they could only detect patterns which the domain experts themselves had identified and were programmed into the system. However, modern machine learning models are able to solve both of these issues. They have the capability to extract their own set of rules which they will base their decisions on and can improve performance over time with additional data. Additionally, the associations which they detect can be ones which experts themselves may be currently unaware of.

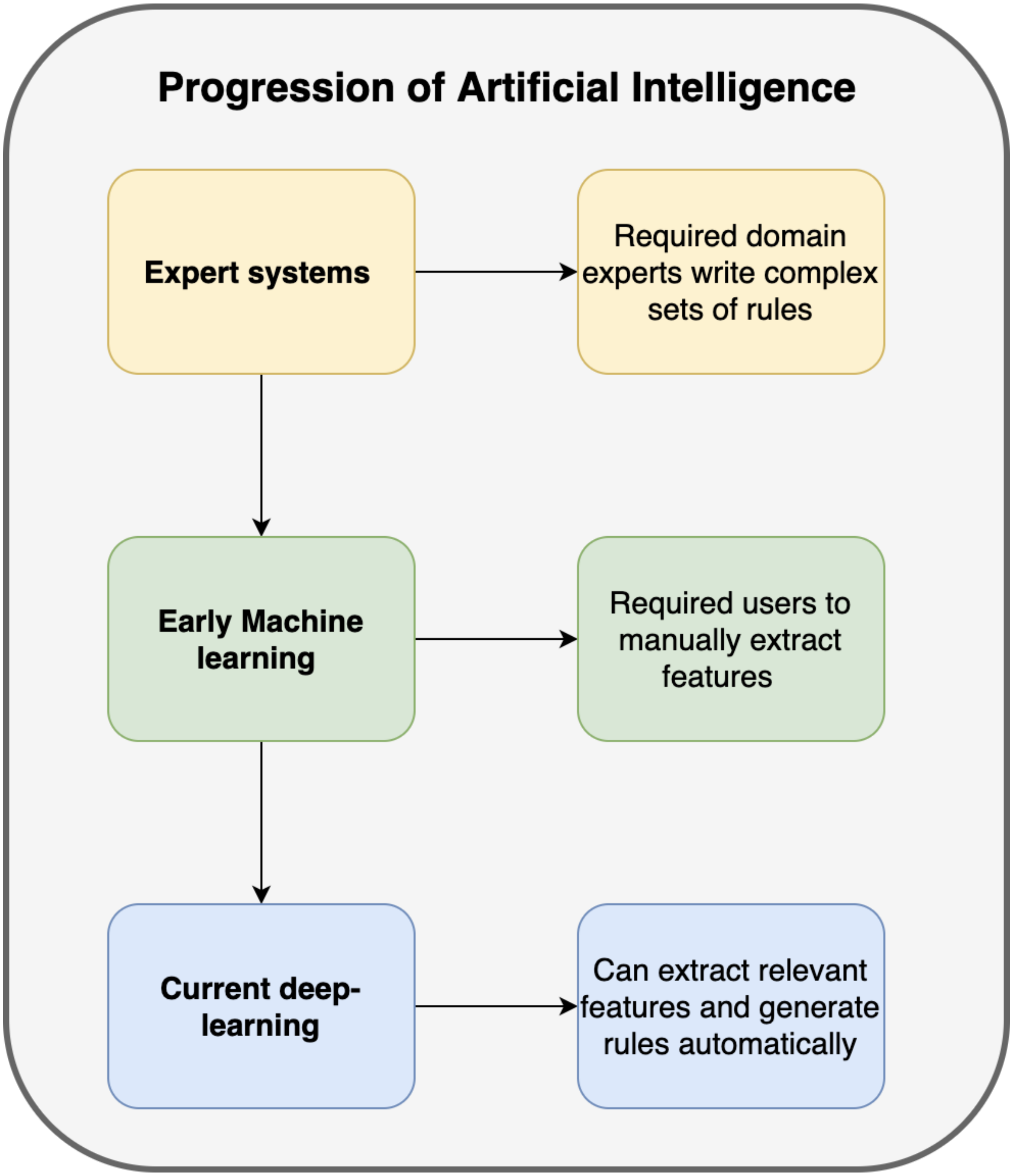

Progression of Artificial Intelligence
Expert systems
Required domain experts write complex sets of rules
Early Machine learning
Required users to manually extract features
Current deep-learning
Can extract relevant features and generate rules automatically

Figure 5. AI is progressing rapidly and increasingly penetrating in medical domain. It already has made significant progress in the field of automated diagnosis.

Before machine learning models can be deployed into the healthcare setting, they must be trained on some data. In the case of the proposed model, it must be trained on different images of breast FNAC slides which have been already labelled by pathologists as either cancerous or non-cancerous. This type of machine learning is called supervised learning where the model is given some input, in this case, an image taken with a microscope of a breast cytologic smear, and there is an expected output associated with each individual image, i.e. is the breast tissue cancerous or non-cancerous. This is known as binary classification where each individual image is assigned to one of two mutually exclusive categories.

Before the machine learning model can be trained, the labelled dataset of images must first be split up into two separate sets: train and test. The training set is what the model is actually trained on. During the process of training, the model is shown an image and it will predict if it is cancerous or non-cancerous. If the model makes an incorrect prediction (i.e. the given image was not cancerous but the model predicted it as cancerous) the model will then update its parameters to make better predictions. This is repeated many times with the goal of selecting optimal parameter values that minimize the error of the model.

The manner in which the data is split between the training set and test set can vary depending on the amount of available data. A common method is to randomly assign elements in the dataset to the training and test set. Common ratios of training and test for small datasets can be around 60% for training and 40% for testing but larger datasets often have smaller relative testing sets (eg: 80% training 20% test). However, a common issue with random sampling is that the ratio of the classes to which the elements belong to may not be consistent between the training and test set. This means that the ratio of cancerous to non-cancerous images in the training set could be significantly different from that of the test set. The method of stratified sampling can ensure balance between the ratio of classes which the elements belong to between the training and test sets (Sechidis, 2011).

In addition to the train and test set, a third set called the validation set can be created. The validation set is used to evaluate the performance of the model on data which it has not learned from during the training phase (Rana 2019). It is important to note though that the model parameters are not updated on images shown in the validation set meaning the model does not learn any features of the images in this set. This is known as cross validation, where the validation set can be used to determine if the model parameters are converging to an optimum value or to prevent overfitting. After the process of training is complete, the model is then evaluated on the test set. This is done to evaluate the real-world performance of the model on data which it has not seen before.

To evaluate the performance of a machine learning diagnostic model, there are a variety of metrics which can be used. Accuracy is a common metric, but it has the issue of being a poor measure of general performance when the dataset is imbalanced (see section on Imbalanced Data). Other metrics such as sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) can provide more a better evaluation of performance than just accuracy alone (Powers, 2011) In the case of the proposed model, the sensitivity shows the ability to correctly diagnose those with breast cancer as having breast cancer and the specificity shows the ability to correctly diagnose those without breast cancer as not having breast cancer.

Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs)

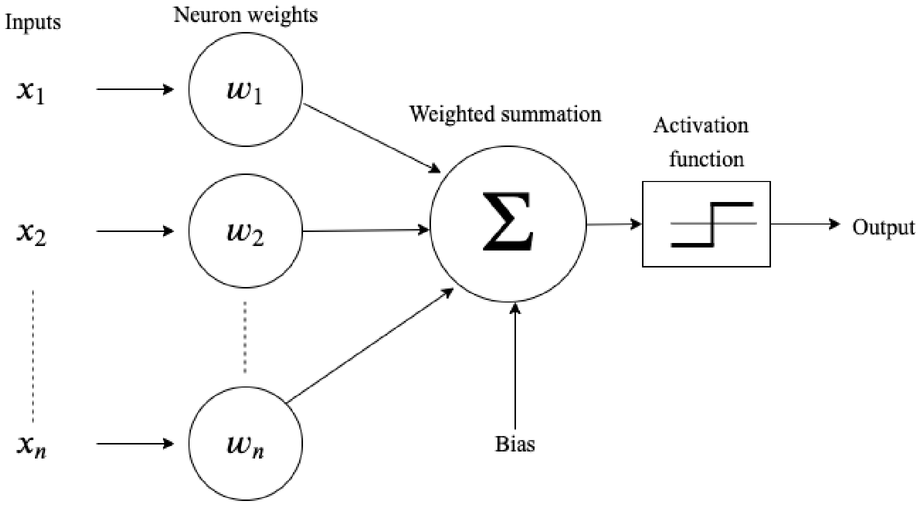


Figure 6: Diagram of a perceptron

Artificial neural networks (ANNs) lie at the core of many of the complex machine learning models which are used today and form the basis for deep learning. Many complex tasks such as image classification or speech recognition employ the use of deep learning. ANNs can be quite robust, are versatile in their applicability and are capable of learning highly complex associations.

Originally taking inspiration from the structure of the brain, ANNs are comprised of units called neurons. An individual neuron is comprised of two parameters which are its weight and bias. An input is received into a neuron and is multiplied by the neuron's weight. Then the bias is added to the result of this and is subsequently passed into an activation function.

A very popular subtype of ANN is a convolutional neural network (CNN). The structure of CNNs take inspiration from the visual cortex of the brain and have been recently shown to be a very powerful tool in computer vision related tasks such as image classification. CNNs are typically comprised of convolutional layers and pooling layers. The first convolutional layer is comprised of neurons which are only connected to a small field of pixels in the input image. Pooling layers typically immediately follow convolutional layers and compress the output from convolutional layers. This is done to reduce the computational complexity of the model by decreasing the number of parameters.

CNNs typically consist of a few convolutional layers followed by a pooling layer and then more convolutional layers followed by a pooling layer and so on. This stacking of convolutional layers followed by pooling layers allows the network to gradually make more abstract representations of characteristics of the image. This means that the first few layers of the network focus on detecting lower level features such as edges and subsequent layers detect higher level features such as recognizing objects.

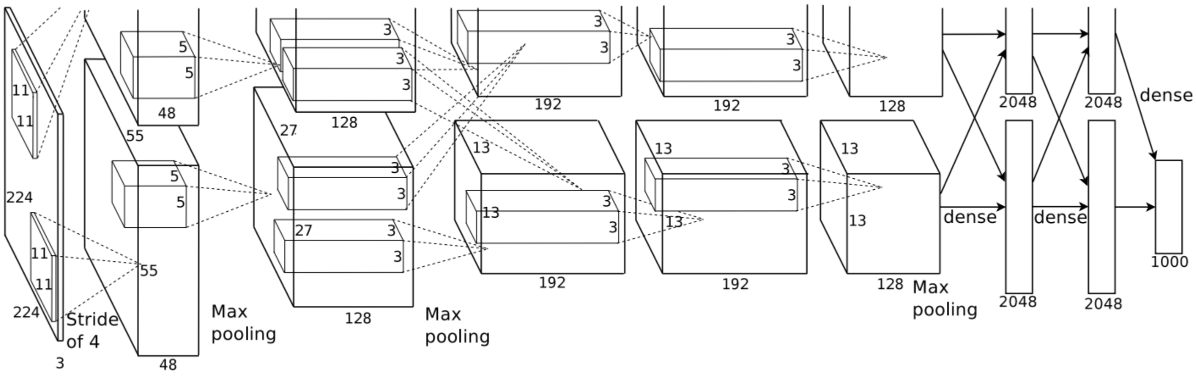


Figure 7: The AlexNet as described in Krizhevsky *et al.* 2012. It was trained on over 1.2 million images belonging to 1000 different categories and has since been used in several medical image classification tasks such as the one in Zejmo *et al.*

Relevant studies

In the paper by Zejmo *et al.*, the CNNs GoogLeNet and AlexNet were used on 50 breast FNAC slides.The methodology involved a pathologist manually extracting 550 regions of interest (ROI) from the slides and the selected ROI were then split into the training and validation set. For each of the 330 ROI in the training set they were split up into 617 different patches of 256x256 pixels. Then, a support vector machine (SVM) classifier, which was manually trained, was used on each patch to determine if there was high enough cellular coverage (e.g. 50% , 85% , 90%) to get an accurate diagnosis. If the patch met the proper cellular coverage threshold, it was then trained by the AlexNet and GoogLeNet CNNs for a fixed number of 10 epochs. The highest performing model which used the GoogLeNet CNN had 86% accuracy.

In the paper by Garud *et al.*, the GoogLeNet CNN was used on 37 breast FNAC slides. A pathologist selected 4-6 high magnification views from each slide which they deemed to contain sufficient cytological evidence to make a diagnosis. Then each high magnification view was pre-processed by applying uniform luminance correction and white balance correction. Then, each view was manually split up into ROI with 918 ROI being created in total. While training, 8 fold cross validation was used on the selected ROI for a fixed number of 200 epochs. A voting scheme classified the image based on the majority prediction of the various ROI and achieved 89.7% mean accuracy.

Conclusions

The issue of breast cancer is one which affects millions around the globe each year with those in developing regions facing disproportionately high rates of fatality. It is known the importance of early diagnosis, but with the relative scarcity and high workload of pathologists in these developing regions, it is vital to research how recent advances in artificial intelligence can serve to aid in the diagnosis of breast cancer. It has been shown that deep learning models already have the capacity for rapid diagnosis with many having the same performance as health-care professionals. In the field of breast cancer, there have been a variety of deep learning diagnostic models created; however, there are very few which have been made to diagnose FNAC results. Since breast cancer diagnosis through FNAC is well suited for developing countries due to its minimal cost and infrastructure requirements, developing a deep learning diagnostic model could be an invaluable tool to assist pathologists working in these regions.

There are a number of challenges associated with building the proposed diagnostic model, but the primary issue is the need for a large, balanced and properly labelled dataset. In the Materials and Methods section, a variety of techniques were discussed in detail to deal with these issues commonly faced with image classification tasks. The majority of these procedures such as regularization to prevent model overfitting, data augmentation to increase diversity of training examples, imputation of missing data entries, oversampling to correct for class imbalance, cost-sensitive learning to minimize false negative diagnosis, semi-supervised learning to handle images with missing labels, and transfer learning are fairly common procedures which many deep learning medical diagnostic models have employed to increase performance.

References

1) Al-Abbadi, M. A. (2011). Basics of cytology. Avicenna journal of medicine, 1(1), 18.

2) Anderson, B. O., Braun, S., Carlson, R. W., Gralow, J. R., Lagios, M. D., Lehman, C., . . . Vargas, H. I. (2003). Overview of breast health care guidelines for countries with limited resources. The breast journal, 9, S42-S50.

3) Antoniou, A., Storkey, A., & Edwards, H. (2017). Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340.

4) Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., . . . Campilho, A. (2017). Classification of breast cancer histology images using convolutional neural networks. PloS one, 12(6).

5) Asia, S., Asia, S., & Hdi, H. Breast Cancer. Source: Globocan 2018. 2018; 876: 2018-2019. In.

6) Becker, A. S., Mueller, M., Stoffel, E., Marcon, M., Ghafoor, S., & Boss, A. (2018). Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. The British journal of radiology, 91(xxxx), 20170576.

7) Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., . . . Balkenhol, M. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama, 318(22), 2199-2210.

8) Byra, M., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., & Andre, M. (2019). Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Medical physics, 46(2), 746-755.

9) Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter, 6(1), 1-6.

10) Coleman, M. P., Quaresma, M., Berrino, F., Lutz, J.-M., De Angelis, R., Capocaccia, R., . . . Hakulinen, T. (2008). Cancer survival in five continents: a worldwide population-based study (CONCORD). The lancet oncology, 9(8), 730-756.

11) Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N. N., . . . Madabhushi, A. (2017). Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. Scientific reports, 7, 46450.

12) Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., . . . Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer, 136(5), E359-E386.

13) Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905.

14) Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321, 321-331.

15) Fujioka, T., Kubota, K., Mori, M., Kikuchi, Y., Katsuta, L., Kasahara, M., . . . Tateishi, U. (2019). Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network. Japanese journal of radiology, 37(6), 466-472.

16) Garud, H., Karri, S. P. K., Sheet, D., Chatterjee, J., Mahadevappa, M., Ray, A. K., . . . Maity, A. K. (2017). High-magnification multi-views based classification of breast fine needle aspiration cytology cell samples using fusion of decisions from deep convolutional networks. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

17) Garud, H., Ray, A. K., Mahadevappa, M., Chatterjee, J., & Mandal, S. (2014). A fast auto white balance scheme for digital pathology. Paper presented at the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI).

18) Gary, M. T., & Tan, P.-H. (2010). Diagnosing breast lesions by fine needle aspiration cytology or core biopsy: which is better? Breast cancer research and treatment, 123(1), 1-8.

19) Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

20) Hukkinen, K., Kivisaari, L., Heikkilä, P. S., Von Smitten, K., & Leidenius, M. (2008). Unsuccessful preoperative biopsies, fine needle aspiration cytology or core needle biopsy, lead to increased costs in the diagnostic workup in breast cancer. Acta Oncologica, 47(6), 1037-1045.

21) Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine, 50(2), 105-115.

22) Khan, M., Haleem, A., Al Hassani, H., & Kfoury, H. (2003). Cytopathological grading, as a predictor of histopathological grade, in ductal carcinoma (NOS) of breast, on air-dried Diff-Quik smears. Diagnostic cytopathology, 29(4), 185-193.

23) Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2017). Deep learning in medical imaging: general overview. Korean journal of radiology, 18(4), 570-584.

24) Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research, 18(1), 559-563.

25) Ling, C. X., & Sheng, V. S. (2008). Cost-sensitive learning and the class imbalance problem. In (Vol. 2011, pp. 231-235): Citeseer.

26) Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., . . . Van Der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports, 6, 26286.

27) Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., . . . Kern, C. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The lancet digital health, 1(6), e271-e297.

28) Łukasiewicz, E., Ziemiecka, A., Jakubowski, W., Vojinovic, J., Bogucevska, M., & Dobruch-Sobczak, K. (2017). Fine-needle versus core-needle biopsy–which one to choose in preoperative assessment of focal lesions in the breasts? Literature review. Journal of ultrasonography, 17(71), 267.

29) Park, Y.-J., Chun, S.-H., & Kim, B.-C. (2011). Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis. Artificial intelligence in medicine, 51(2), 133-145.

30) Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

31) Pigott, T. D. (2001). A review of methods for missing data. Educational research and evaluation, 7(4), 353-383.

32) Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

33) Rana, S. P., Dey, M., Tiberi, G., Sani, L., Vispa, A., Raspa, G., . . . Dudley, S. (2019). Machine Learning Approaches for Automated Lesion Detection in Microwave Breast Imaging Clinical Data. Scientific reports, 9(1), 1-12.

34) Ross Michael, H. (2010). Histology: a text and atlas: with correlated cell and molecular biology/Michael H. Ross, Wojciech Pawlina. In: Philadelphia: Wolters Kluwer Health,-2016.–984 p.

35) Ruifrok, A. C., & Johnston, D. A. (2001). Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology, 23(4), 291-299.

36) Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

37) Smith, R. A., Caleffi, M., Albert, U. S., Chen, T. H., Duffy, S. W., Franceschi, D., . . . Panel, A. t. C. (2006). Breast cancer in limited-resource countries: early detection and access to care. The breast journal, 12, S16-S26.

38) Teramoto, A., Tsukamoto, T., Kiriyama, Y., & Fujita, H. (2017). Automated classification of lung cancer types from cytological images using deep convolutional neural networks. BioMed research international, 2017.

39) Thai-Nghe, N., Gantner, Z., & Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. Paper presented at the The 2010 International joint conference on neural networks (IJCNN).

40) Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44-56.

41) Torrey, L., & Shavlik, J. (2010). Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques (pp. 242-264): IGI Global.

42) Unger-Saldaña, K. (2014). Challenges to the early diagnosis and treatment of breast cancer in developing countries. World journal of clinical oncology, 5(3), 465.

43) Vimpeli, S.-M., Saarenmaa, I., Huhtala, H., & Soimakallio, S. (2008). Large-core needle biopsy versus fine-needle aspiration biopsy in solid breast lesions: comparison of costs and diagnostic value. Acta Radiologica, 49(8), 863-869.

44) Willems, S. M., Van Deurzen, C., & Van Diest, P. (2012). Diagnosis of breast lesions: fine-needle aspiration cytology or core needle biopsy? A review. Journal of clinical pathology, 65(4), 287-292.

45) Yamaguchi, R., Tsuchiya, S.-I ., Koshikawa, T., Ishihara, A., Masuda, S., Maeda, I., . . . Narita, M. (2012). Diagnostic accuracy of fine-needle aspiration cytology of the breast in Japan: report from the Working Group on the Accuracy of Breast Fine-Needle Aspiration Cytology of the Japanese Society of Clinical Cytology. Oncology reports, 28(5), 1606-1612.

46) Young, I. T. (2000). Shading correction: compensation for illumination and sensor inhomogeneities. Current Protocols in Cytometry, 14(1), 2.11. 11-12.11. 12.

47) Żejmo, M., Kowal, M., Korbicz, J., & Monczak, R. (2017). Classification of breast cancer cytological specimen using convolutional neural network. Paper presented at the Journal of Physics: Conference Series.