UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Khatia Kilanava

Enhancing Breast Cancer Prediction Using Unlabeled Data

Master's Thesis (30 ECTS)

Supervisor: Prof. Sherif Aly Ahmed Sakr
Supervisor: Dr. Radwa El Shawi

Tartu 2019

Enhancing Breast Cancer Prediction Using Unlabeled Data

Abstract: The following thesis presents a deep learning (DL) approach for automatic classification of invasive ductal carcinoma (IDC) tissue regions in whole slide images (WSI) of breast cancer (BC) using unlabeled data. DL methods are similar to the way the human brain works across different interpretation levels. These techniques have shown to outperform traditional approaches of the most complex problems such as image classification [KSH12] and object detection [AAL+15]. However, DL requires a broad set of labeled data that is difficult to obtain, especially in the medical field as neither the hospitals nor the patients are willing to reveal such sensitive information. Moreover, machine learning (ML) systems are achieving better performance at the cost of becoming increasingly complex. Because of that, they become less interpretable that causes distrust from the users. Model interpretability is a way to enhance trust in a system. It is a very desirable property, especially crucial with the pervasive adoption of ML-based models in the critical domains like the medical field. With medical diagnostics, the predictions cannot be blindly followed as it may result in harm to the patient. IDC is one of the most common and aggressive subtypes of all breast cancers accounting nearly 80% [DSBJ11] of them. Assessment of the disease is a very time-consuming and challenging task for pathologists, as it involves scanning large swatches of benign regions to identify an area of malignancy. Meanwhile, accurate delineation of IDC in WSI is crucial for the estimation of grading cancer aggressiveness. In the following study, a semi-supervised learning (SSL) scheme is developed using the deep convolutional neural network (CNN) for IDC diagnosis. The proposed framework first augments a small set of labeled data with synthetic medical images, generated by the generative adversarial network (GAN) that is followed by feature extraction using already pre-trained network on the larger dataset and a data labeling algorithm that labels a much broader set of unlabeled data. After feeding the newly labeled set into the proposed CNN model, acceptable performance is achieved: the AUC and the F-measure accounting for 0.86, 0.77, respectively. Moreover, proposed interpretability techniques produce explanations for medical predictions and build trust in the presented CNN. The following study demonstrates that it is possible to enable a better understanding of the CNN decisions by visualizing areas that are the most important for a particular prediction and by finding elements that are the reasons for IDC, Non-IDC decisions made by the network.

Keywords:
Invasive ductal carcinoma, interpretability, convolutional neural networks, whole-slide imaging, generative adversarial networks, machine learning

CERCS: P170 Computer science, numerical analysis, systems, control

Sildistamata andmete kasutamine rinnavähi ennustamise parendamiseks

Lühikokkuvõte: Selles väitekirjas esitatakse sildistamata andmeid kasutav süvaõppe lähenemine rinna infiltratiivse duktaalse kartsinoomi koeregioonide automaatseks klassifitseerimiseks rinnavähi patoloogilistes digipreparaatides. Süvaõppe meetodite tööpõhimõte on sarnane inimajule, mis töötab samuti mitmetel tõlgendustasanditel. Need meetodid on osutunud tulemuslikeks ka väga keerukate probleemide nagu pildiliigituse [KSH12] ja esemetuvastuse lahendamisel [AAL+15], ületades seejuures varasemate lahendusviiside efektiivsust. Süvaõppeks on aga vaja suurt hulka sildistatud andmeid, mida võib olla keeruline saada, eriti veel meditsiinis, kuna nii haiglad kui ka patsiendid ei pruugi olla nõus sedavõrd delikaatset teavet loovutama. Lisaks sellele on masinõppesüsteemide saavutatavate aina paremate tulemuste hinnaks nende süsteemide sisemise keerukuse kasv. Selle sisemise keerukuse tõttu muutub raskemaks ka nende süsteemide töö mõistmine, mistõttu kasutajad ei kipu neid usaldama. Meditsiinilisi diagnoose ei saa järgida pimesi, kuna see võib endaga kaasa tuua patsiendi tervise kahjustamise. Mudeli mõistetavuse tagamine on seega oluline viis süsteemi usaldatavuse tõstmiseks, eriti just masinõppel põhinevate mudelite laialdasel rakendamisel sellistel kriitilise tähtsusega aladel nagu seda on meditsiin. Infiltratiivne duktaalne kartsinoom on üks levinumaid ja ka agressiivsemaid rinnavähi vorme, moodustades peaaegu 80% [DSBJ11] kõigist juhtumitest. Selle diagnoosimine on patoloogidele väga keerukas ja ajakulukas ülesanne, kuna nõuab võimalike pahaloomuliste kasvajate avastamiseks paljude healoomuliste piirkondade uurimist. Samas on infiltratiivse duktaalse kartsinoomi digipatoloogias täpne piiritlemine vähi agressiivsuse hindamise aspektist ülimalt oluline. Käesolevas uurimuses kasutatakse konvolutsioonilist närvivõrku arendamaks välja infiltratiivse duktaalse kartsinoomi diagnoosimisel rakendatav pooleldi juhitud õppe skeem. Välja pakutud raamistik suurendab esmalt väikest sildistatud andmete hulka generatiivse võistlusliku võrgu loodud sünteetiliste meditsiiniliste kujutistega. Seejärel kasutatakse juba eelnevalt treenitud võrku, et selle suurendatud andmekogumi peal läbi viia kujutuvastus, misjärel sildistamata andmed sildistatakse andmesildistusalgoritmiga. Töötluse tulemusena saadud sildistatud andmeid eelmainitud konvolutsioonilisse närvivõrku sisestades saavutatakse rahuldav tulemus: ROC kõvera alla jääv pindala ja F1 skoor on vastavalt 0,86 ja 0,77. Lisaks sellele võimaldavad välja pakutud mõistetavuse tõstmise tehnikad näha ka meditsiinilistele prognooside otsuse tegemise protsessi seletust, mis omakorda teeb süsteemi usaldamise kasutajatele lihtsamaks. Käesolev uurimus näitab, et konvolutsioonilise närvivõrgu tehtud otsuseid aitab paremini mõista see, kui kasutajatele visualiseeritakse konkreetse juhtumi puhul infiltratiivse duktaalse kartsinoomi positiivse või negatiivse otsuse langetamisel süsteemi jaoks kõige olulisemaks osutunud piirkondi.

Märksõnad:
infiltratiivne duktaalne kartsinoom, konvolutsiooniline närvivõrk, digipatoloogia, digi-

taalpatoloogia, generatiivne võistluslik võrk, masinõpe

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Ackowledgement

I want to thank the Data Systems Group for having me on their team, allowing me to work on such an exciting topic, and for enabling me to use all the necessary resources that were demanded by the project. A special thanks to my supervisors Prof. Sherif Aly Ahmed Sakr and Dr. Radwa El Shawi for suggesting the topic, and for helping me throughout this period through their advice and feedback. It was fascinating to see how computer science can be applied to solve issues in the medical field.

Abbreviations

ADN	Adaptive Reconvolutional Network
CNN	Convolutional Neural Networks
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
DP	Digital Pathodology
DPI	Digital Pathodology images
GAN	Generative Adversarial Network
GB	Gradient Boost
NAS	Neural Architecture Search
SAE	Stacked Auto-Encoders
SSL	Semi Supervised Learning
ReLU	Rectified Liner Unit
RBM	Restricted Boltzann Machines
RF	Random Forrest
WSI	Whole Slide Imaging

Contents

Abstract **2**

Ackowledgement **5**

Abbreviations **6**

1 Introduction **9**
1.1 Contribution 9
1.2 Thesis outline 10

2 Related work **11**
2.1 Deep learning applications in medical analysis 11
2.1.1 Supervised deep learning 11
2.1.2 Weakly supervised deep learning 13
2.1.3 Unsupervised deep learning 14
2.2 Neural Architecture Search 15
2.2.1 Applications in NAS 16
2.2.2 Neural Network Intelligence 16
2.3 Augmentation techniques 17
2.3.1 Manual data augmentation 17
2.3.2 Augmentation with GANs 18
2.4 Feature extraction techniques 19
2.4.1 Handcrafted features 20
2.4.2 Transfer learning 20
2.5 Interpretability 21
2.5.1 Importance of interpretability 21
2.5.2 Local model interpretability 22
2.5.3 Global model Iterpretability 23

3 Methodology **24**
3.1 Proposed framework 24
3.2 Dataset 24
3.3 Baseline 25
3.4 Data augmentation 27
3.5 Feature extraction 27
3.6 Data labeling 27
3.7 Neural network 28
3.8 Model interpretability 29

3.8.1 Grad-Cam 29
3.8.2 Regression Concept Vectors 30
3.8.2.1 Framework 30
3.8.2.2 Sensitivity scores 32
3.8.2.3 Bidirectional Relevance Score 33
3.9 Experimental setup 33

4 Results **35**
4.1 Experimental design 35
4.2 Performance measures 35
4.3 Data Labeling method 35
4.4 CNN model performance 36
4.5 Model interpretability 37
4.6 Discussion 38

5 Conclusion **41**

References **56**

Appendix **57**
II. Licence 57

1 Introduction

IDC is one of the most prevalent and dangerous breast cancers in women, and identification of its aggressiveness during the early stages is the most crucial step in treatment. In most cases, the survival rate is dependent on early diagnosis. However, it is a very challenging task for pathologists to identify the malignant area among huge swatches of benign regions. In supervised learning, the performance of a model is dependent on the amount of labeled data. The greater the volume of the training set available for the classifier, the more it can learn. However, it is a very time-consuming, expensive, and difficult task to collect enough labeled data, especially medical data like annotation of IDC areas on the WSI. It is also very challenging to extract handcrafted features from medical images, as many factors are still restricting the efficacy of screening mammography, and many elements vary from one medical case to another. Neural Networks, on the other hand, solve the problem of feature engineering and have shown enormous potential as a successful application in machine learning; from image classification [KSH12] to object detection [GDDM14], image captioning [VTBE15], visual question answering [AAL+15] and many other domains. However, this is highly dependent on the number of labeled instances, which can be difficult to obtain. To label large sets of examples, several radiologists are needed first to label the data, and then compare the results to one another case by case. One problem is the workload, and another is a consensus between results as many conflicts can arise. Besides, it is not possible to get these massive labeled datasets from hospitals and similar institutions since neither the doctors nor the patients are willing to reveal that information due to privacy or policy of the particular institution. Model interpretability matters. To trust intelligent systems like Deep Neural Networks (DNNs) and integrate them with everyday life, they have to be transparent. It has to be completely clear why the model makes a particular decision. In general, transparency is essential for three reasons. First, when AI is weaker than humans, the goal is to identify where it fails to help researchers to focus more on improving those aspects [ABP16]. Second, when AI is as powerful as human, users need enough confidence and trust to rely on it. Finally, when AI is much more potent than humans [SHM+16], the goal is for users to learn from the machines how to make better decisions. Decision-making based on a false premise is especially highly undesirable in applications like medical imaging as it can easily lead to someone's death. Models, like DNNs, cannot be integrated into the medical domain, until they are considered as "black box" models.

1.1 Contribution

In this work, the semi-supervised framework for breast cancer diagnosis is proposed. The problems related to the lack of sufficient labeled data and model interpretability are addressed. It will be shown that it is possible to label the significantly large amount of

unlabeled sets successfully. Using a small portion of labeled data, the model achieves almost as good performance as if the model was trained on the larger set of labeled instances. Further, when training on the fully labeled instances, the proposed CNN outperforms other papers working on the same dataset [CRBG$^+$14][JM16]. Finally, the model interpretability issue is addressed by explaining model predictions on the test set both locally and globally.

1.2 Thesis outline

In Section 2, the state-of-the-art techniques applied in both supervised and semi-supervised DL are presented. Moreover, augmentation and feature extraction methods are also reviewed. The section ends with a general overview of model interpretability techniques. In Section 3, the detailed framework of the study is presented. Section 4 shows different results and discusses them. The conclusion and future prospects can be found in Section 5.

2 Related work

This section reviews the work related to automated medical image analysis using DL, various feature extraction, feature augmentation techniques, state-of-the-art model interpretability methods, and identifies their possible limitations.

2.1 Deep learning applications in medical analysis

Related work to applications of DL can be categorized into supervised, unsupervised, and weakly supervised approaches. This section considers the work applied on only the medical domains and reviews detection, segmentation, and classification techniques sorted by different medical modalities, including breast, chest, lung, and many other abnormalities. The section is based on the information extracted from the Litjens et al. survey paper [LKB+17] on DL in medical images analysis.

2.1.1 Supervised deep learning

Researchers started developing the models for automated medical analysis as soon as it became available to transfer medical images to a computer for further processing. Thus, by the end of the 1990s, the interest and popularity of the supervised learning system development started to increase. The most critical and time-consuming step while using such models is feature engineering when extensive information is extracted from the images, that are called handcrafted features. By using DL models, higher dimensional features are learned. These models consist of several layers that gradually transform the input images to the output that in a medical scenario is the presence or absence of the disease of the particular person. In the medical domain, it was first applied by Lo et al. [LLL+95].
DL contribution towards the medical image classification is significant. In the typical scenario, the model can output a target variable given the input images.
Chest x-ray image analysis started from early 1995, by Lo et al. [LLL+95] who detected malignant nodules using a two-layer CNN. There are several papers published in the years of 2015-2016 in this area. An image retrieval application was made by Anavi et al. [AKG+16] who obtained the features from the last fully connected layer of a five-layer CNN that were feed to a one-vs-all support vector machine (SVM) for classification. Some of the works for chest x-ray analysis are summarised in Table 1.
There were several works for pathology detection on x-ray images. Bar et al. [BDW+18] extracted low-level features in order to detect several different diseases, whereas Hwang et al. [HK16] detected Tuberculosis still using CNN but fine-tuned on the higher volume datasets. Rajkomar et al. [RLT+17] and Wang et al. [WKG+16] also used ImageNet pre-trained network for frontal/lateral, and nodule classification but Wang et al. used the network for feature retrieval and combined it with the original features, whereas

Table 1. Overview of papers using deep learning techniques for chest x-ray analysis

Reference	Method	Application
Lo et al. (1995)	Nodule detection	Classifies candidates from small patches with a two-layer CNN
Anavi et al.(2015)	Image retrieval	Combines classical features with those from a pre-trained CNN for image retrieval
Bar et al. (2015)	Pathology detection	Features from a pre-trained CNN and low-level features are used to detect various diseases
Bar et al. (2016)	Image retrieval	Continuation of Anavi et al. (2015), adding age and gender as features
Bar et al. (2016)	Pathology detection	Continuation of Bar et al. (2015), more experiments and adding feature selection
Hwang et al. (2016)	Tuberculosis detection	Processes entire radiographs with a pre-trained, fine-tuned network with 6 convolution layers
Rajkomar et al. (2017).	Frontal/lateral classification	A pre-trained CNN performs frontal/lateral classification task
Wang et al. (2016)	Nodule classification	Combines classical features with CNN features from a pre-trained CNN on ImageNet

Rajkomar et al. performed standard classification task.
Since this study works on mammography (MG) images, it worth mentioning other related papers within this domain. Table 2 gives a summary of related papers. The oldest research in this area is done by Sahiner et al. [SCP+96] using CNN as a method for classification. In 2012 Jamieson et al. [JDG12] implemented four layer Adaptive Reconvolutional Network (ADN) that is an older version of a CNN. Use of pre-trained CNN on natural image patches is a common technique on MG images as well. Huynh et al. [HLG16], Samala et al. [SCH+16], Kooi et al. [KvGKdH17], Fonseca et al. [FMW+15] all used pre-trained network either for feature-extraction tool followed by SVM or for direct classification of new images.
Digital Pathology (DP) is a part of virtual microscopy, and it is a process in which scientists convert glass slides into digital slides, that are then transferred to a computer, preprocessed and analyzed digitally. This way it is much easier to view, analyze and process the information and with the help of WSI, this area gives even more promising results as it achieves faster performance and affordable diagnosis and prognosis various important diseases. Table 3 gives an overview of the papers using DP for digital pathology images (DPI). There are various studies performed for either detecting or classifying the disease on the DPI. Kim et al. [KCRB16] used AlexNet architecture that was further fine-tuned on the cytology images for thyroid cytopathology classification. Authors show that the method can leverage networks that have been trained on an extensive set of generic images, to medical scenarios where only a hundred of images is available. A

Table 2. Overview of papers using deep learning for breast image analysis

Reference	Method	Application
Sahiner et al. (1996)	CNN	First application of a CNN to mammography
Jamieson et al. (2012)	ADN	Four layer ADN, an early form of CNN for mass classification
Fronseca et al. (2015)	CNN	Pre-trained network extracted features classified with SVM for breast density estimation
Huynh et al. (2016)	CNN	Pre-trained CNN on natural image patches applied to mass classification
Samala et al. (2016)	CNN	Pre-trained CNN on mammographic masses transfered to tomosynthesis
kooi et al. (2017)	CNN	Pre-trained CNN on mass/normal patches to discriminate malignant masses from (benign) cysts.

Table 3. Overview of papers using deep learning for digital pathology images

Reference	Topic	Method
Cruz-Roa et al. (2014)	Detecting of invasive ductal carcinoma	CNN-based patch classifier
Xu et al. (2014)	Patch-level classification of colon cancer	Multiple instance learning framework with CNN features
Bychkov et al. (2016)	Outcome prediction of colorectal cancer	Extracted CNN features from epithelial tissue for prediction
Kallen et al. (2016)	Predicting Gleason score	Overfeat pre-trained network as feature extractor
Kim et al. (2016)	Thyroid cytopathology classification	Fine-tuning pre-trained AlexNet
Schaumberg et al. (2016)	SPOP mutation prediction of prostate cancer	Ensebmle ResNets
Wng et al. (2016).	Metastases detection in lymph node	Ensemble of CNNs with hard negative mining

couple of papers – Schaumberg et al. [SRF16] and Wang et al. [WKG+16], proposed an ensemble of ResNet and CNN respectively for prostate cancer classification and lymph node detection. Kallen et al. [KMH+16] used the Overfeat network for feature extraction, which is very rarely used in medical images.

2.1.2 Weakly supervised deep learning

Models learn more when they have enough labeled data available, which results in a better performance. However, collecting enough labeled set is tricky as it requires time and enough resources. SSL makes use of unlabeled and labeled data by making a cluster assumption: similar attributes lead to similar labels. SSL might be useful in the cases when labeled data is not enough to get satisfactory results. Zhu et al. [ZLR05], Shin et al. [STS+06], Shin et al. [SLL07] have even shown that learning with a mixed set of labeled and unlabeled data can achieve better performance than having labeled data only. As medical datasets are growing in size, there is always a lack of annotated instances. As a result, SSL is actively used in medical scenarios. This section reviews some of the cases of SSL both in diagnosis/detection and segmentation tasks. The information is extracted from the Cheplygina et al. [CdBP19] survey paper.

The types of SSL are self-training and co-training. In both cases, a learner is trained on the labeled instances first, and it is used to label the rest of the unlabeled set. The newly labeled set is then appended in the originally labeled data. Many papers apply active learning, and experts are asked to verify the labels (Parag et al. [PPS14], Su et al. [SYH+16]).

Self-training is mainly used for segmentation in various medical cases: in the brain by Iglesias et al. [ILTT10], Meier et al. [MBS+14], Wang et al. [WLP+14], Dittrich et al. [DRK+14], in the retina by Gu et al. [GZB+17], in the histology and microscopy by Singh et al. [SJP+11] for nuclei cell classification.

Another type of SSL is Graph-based method. In that case, graphs are constructed with samples as nodes and edges as similarities between samples. This type of SSL makes a different assumption. It is assumed that similar samples have a similar distance from the neighboring nodes. For segmentation, it was used by Gass et al. [GSG12], Song

Table 4. Overview of semi-supervised learning applications.

Reference	Application	SSL category
	Brain	
Song et al. (2009)	tumor segmentation	graph-based
Iglesias et al. (2010)	skull stripping	self-training
Filipovych et al. (2011)	Classification if MCI	semi-supervised SVM
Batmanghelich et al. (2011)	Classification of AD, MCI	graph-based
Meier et al. (2014)	tumor segmentation	graph based
Dittrich et al. (2014)	fetal brain segmentation	self-training
Wang et al. (2014).	lesion segmentation	self-training
An et al. (2016)	AD classification	graph-based
Moradi et al. (2015)	classification of MCI	semi-supervised SVM
	Histology and microscopy	
Singh et al. (2011)	cell type classification in microscopy	self-training
Parag et al. (2014)	cell type segmentation in microscopy	graph-based active
Su et al. (2016)	cell segmentation in microscopy	graph-based, active
	Multiple	
Gass et al. (2012)	segmentation in two applications	graph-based
Ciurte et al. (2014)	segmentation in four applications	graph-based
Gu et al. (2017)	segmentation in two applications	self-training

et al. [SZL+09] and Ciurte et al. [CBC+]. Sun et al. [STZQ16] identified malignant and benign cases of breast cancer with the help of CNN, but before the neural network, SSL graph-based method was used to label much larger unlabeled sets using minimal training data. The study showed that unlabeled data provides some additional valuable information for the model.
Manifold regularization is similar to the SSL graph-based method and can generalize unseen labels. Such an approach is not only used for segmentation purposes but computer-aided diagnosis as well. The works by An et al. [AAL+16] and Batmanghelich et al. [BDP+11] demonstrated this.
Semi-supervised SVMs are based on the assumption that there should be a low-density region along with different classes. It is used for classification of AD and MCI (Filipovych et al. [FDI+11], Moradi et al. [MPG+15]). For a general overview of different applications of SSL, refer to Table 4.

2.1.3 Unsupervised deep learning

Unsupervised algorithms process data without labels, thus models are trained to find patterns, like latent subspaces. The method gives the possibility to make use of all unlabeled data available in the world. Traditional unsupervised models are PCA and clustering methods. Recently, the community has been focused on unsupervised pre-training and network architectures like stacked auto-encoders (SAEs) and Restricted Boltzmann machines (RBMs).
Uzunova et al. [UHE18] proposed two different unsupervised detection approaches: 1) The PatchMatch algorithm which compares healthy patient images, to the ones that

have a pathology. 2) Conditional variational auto-encoders (CVAE), which model healthy data by mapping it to the lower dimensional latent space. Kallenberg et al. [KPN+16] proposed Unsupervised CNN feature learning with SAE for breast density classification. Various research has been done in DL for brain image analysis. Table

Table 5. Overview of papers using deep learning techniques for brain image analysis. All works use MRI unless otherwise mentioned

Reference	Method	Application
Brosch and Tam (2013)	DBN	AD/HC classification; Deep belief networks with convolutional RBMs for manifold learning
Plis et al.	DBN	Deep belief networks evaluated on brain network estimation, Schizophrenia and Huntington's disease classification
Suk and Shen (2013)	SAE	AD/MCI classification; Stacked auto encoders with supervised fine-tuning
Suk et al. (2014)	RBM	AD/MCI/HC classification; Deep Boltzmann Machines on MRI and PET modalities
Payan and Montana (2015)	CNN	AD/MCI/HC classification; 3D CNN pre-trained with sparse auto-encoders
Hossein-Asl et al. (2016)	CNN	AD/MCI/HC classification; 3D CNN pre-trained with a 3D convolutional auto-encoder on fMRI data
Kim et al.	ANN	Schizophrenia/NH classification on fMRI; Neural network showing the advantage of pre-training with SAEs, and L1 sparsification

5 gives an overview of the papers performing brain disorder classification; most of them work on MRI. Brosch et al. [BTI+13] classified Alzheimer's disease using deep belief networks and convolution RBMs for manifold learning. Payan et al. [PM15] and Hosseini-asl et al. [HAGEB16] were working on the same disease but used 3D CNNs instead of 2D. Specifically, they used SAEs alongside 3D CNNs to build a model that can differentiate between the disease cases of a patient, using an MRI scan of a brain. Three other papers were studying on Schizophrenia- Suk et al. [SLS+14], Suk et al. [SS13], Plis et al. [PHS+14], and all were working on the MRI scans and used DBNs, RBMs and SAEs respectively. Kim et al. [KCRB16] used fMRI scans instead of MRI and showed the positive outcomes of using a pre-training network with SAEs. Another unsupervised learning examples are GANs. GANs are a special type of neural network models where two networks are trained simultaneously. More details of GAN architecture will be reviewed in Section 2.3.2. GANs are mostly used as a data augmentation technique. However, there are several studies, that used it for unsupervised classification and detection tasks - Hu et al. [HTC+17] combined WGAN and InfroGAN for unsupervised learning in histopathology images. Schleg et al. [SSW+17] used GAN to learn a manifold of normal anatomical variability and presented a state-of-the-art anomaly scoring scheme. Alex et al. [AKCK17] used GAN for brain lesion detection. The generator was used to learn the distribution of regular patches, whereas discriminator was used to compute a posterior probability of every test image.

2.2 Neural Architecture Search

The success of DL is mainly due to its automation of the feature engineering process. However, neural architecture engineering is a challenging and time-consuming step. Currently, employed architectures are mostly developed by human experts and might be

an error-prone process. Because of this, there is a rising interest in making neural architecture search (NAS) automated. The goal of Automated Machine Learning (AutoML) is to enable people without ML background to use ML models easily. NAS aims to search for the best architectural design and the best parameter set for the given learning task and dataset in a reasonable time. The following section gives an overview of existing applications in Neural Architecture Search (NAS).

2.2.1 Applications in NAS

Early NAS works tried to search for a complete Network [LLS+17], while more recent ones tried finding healthy cells [LZN+18]. A notable drawback of the mentioned approaches is expensive computational overhead. DARTS (Differentiable Architecture Search), introduced by [LSY18] achieved state-of-the-art results and a remarkable performance by getting rid of the time-consuming process of architecture evaluating and sampling. The main idea in DARTS is that search space is continuous instead of being discrete that avoids searching in a broad set of candidate architectures. A general overview of DARTS framework is as follows: operation on edges (connection of 2 neurons) is initially unknown, and a continuous relaxation of search space is followed by referring to a mixture of candidate operations per each edge. Finally, bilevel optimization is solved to jointly optimize mixing probabilities and network weights. The final architecture is obtained from inducing the learned probabilities. Another relevant topic is One-Shot [BKZ+18] that trains the over-parameterized network with DropPath [ZVSL18] and gets rid of each path with some probability. Then, the pre-trained system is used to evaluate architectures that are sampled from zeroing out paths. However, this type of NAS suffers from large GPU memory consumption. Stochastic neural architecture search (SNAS) proposed by [XZLL18] follows the pipeline of DARTS, and the main idea is to build a very efficient end-to-end system with the smallest NAS framework possible. This is achieved by forcing the weights on each edge to be one-hot, that tackles the inconsistency between search and evaluation optimization. SNAS trains neural operation parameters as well as architecture distribution in the same round of back-propagation. Another novel approach is ProxylessNAS (direct neural architecture search) [CZH18], which learns architecture on the target task without a proxy dataset. The technique achieves memory efficiency by adopting binary masks to operations and by allowing only one path of the parameterized network to be activated and fed in the GPU. For a General overview of applications in NAS, refer to Table 6.

2.2.2 Neural Network Intelligence

For the best architecture search and tuning this study applies Neural Network Intelligence (NNI) proposed by Microsoft. NNI is an online platform which enables users to find the best achieving neural network architectures of their particular dataset. Besides, the

Table 6. Overview of applications in NAS

Method	Remarks	Refference
DARTS	memory efficient- gets rid of architecture evaluation and sampling	[LSY18]
SNAS	The smallest NAS framework	[BKZ+18]
One-Shot	Large GPU consumption	[XZLL18]
ProxyLessNAS	Memory efficient	[CZH18]

toolkit helps to tune machine learning models by finding a set of hyperparameters. The general process works as follows: a tuner, which is an AutoML algorithm, receives a search space predefined by users and generates configurations. Then an iterative process is started: configurations are submitted for training, and performance is sent back to the tuner which creates new configurations and provides for training again. For one particular experiment, users only have to define a search space that is a space for tuning the model like the value ranges for hyperparameters, update model codes, and run the experiment. The results can be obtained within each trial. Figure 1 shows the structure of the NNI.
NNI provides recent builtin-tuners that are quite easy to use. This study uses a Tree-structured Parzen Estimator (TPE) tuner for hyperparameter search. TPE is a sequential model-based optimization tool that is used in various scenarios and shows excellent performance. TPE is mainly suggested in case of limited computation resources and authors claim that it is far better than Random Search.
For neural architecture search, the tuner Network Morphism is used. It provides the abilities to search for the best performing architecture automatically. Every child of a network gets the knowledge from its parent and morphs into different types of networks. The value of a child network is estimated using historical architecture. After this, the best one is chosen to be trained.
There are some other well-known tuners available to choose from like Random Search-surprisingly simple and effective search and Grid Search – more exhaustive searching in manually specified hyperparameters space.

2.3 Augmentation techniques

This section gives a general overview of the existing manual data augmentation techniques and presents the studies, which used GAN as a powerful tool for synthetic image synthesis on a different medical domain.

2.3.1 Manual data augmentation

Various methods have been proposed as an augmentation tool, including horizontal flips, random crops, and color jittering [SRASC14] on different modalities. Krichevsky et al.

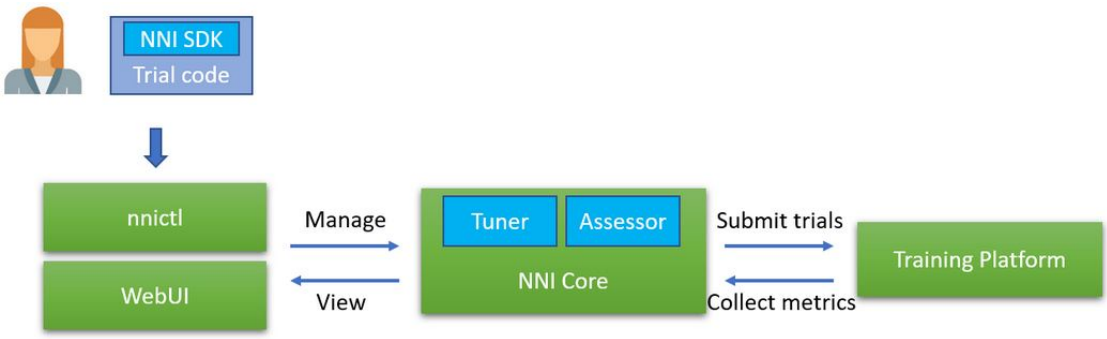


Figure 1. NNI architecture [Win]

applied a technique fancy PCA that alters the intensities of the RGB channels on the images [KSH12]. Hussain et al. [HGYR17] experimented with different augmentation techniques to find out which one leads on more discriminative models and concluded that flips and Gaussian filters achieve the best performance whereas adding noise leads to poor results.
Other common forms of data augmentations are simple transformations like left-right flipping, translations, rotations, scaling of images [KSH12][CB16]. Flipping, resizing, cropping are also widely used [HB04]. Besides, geometric distortions of deformations are used to add several samples in the training set [KMG17a], or to balance the size of the dataset [KMG17b]. Popular methods are histogram equalization, white-balancing, blurring, sharpening. Those methods are fast, reproducible, and relatively easy for code implementation.

2.3.2 Augmentation with GANs

GAN is a new and more powerful tool to perform synthetic image generation. GANs utilize two different networks – a discriminator D(x) and a generator G(z). It is a minimax game were generator seeks to generate as realistic pictures as possible to fool the discriminator, whereas discriminator is trying to accurately distinguish between fake and generated images as shown on Figure 2.
Unconditional synthesis is a process of image generation from a random noise without any conditions. Techniques commonly used for medical image synthesis are summarized in Table 7 that is obtained from Yi et al. [YWB18] survey paper about GAN applications in medical imaging. The mostly used and most successful techniques in medical imaging are DCGAN [RMC15], WGAN [GAA$^+$17] and PGGAN [KALL17] because of their good training stability and performance. Most studies train a separate generator for each class. For instance, Frid-Adar et al. [FADK$^+$18] used three DCGANs to obtain fake images from all the three classes. This work applies DCGAN for synthetic image

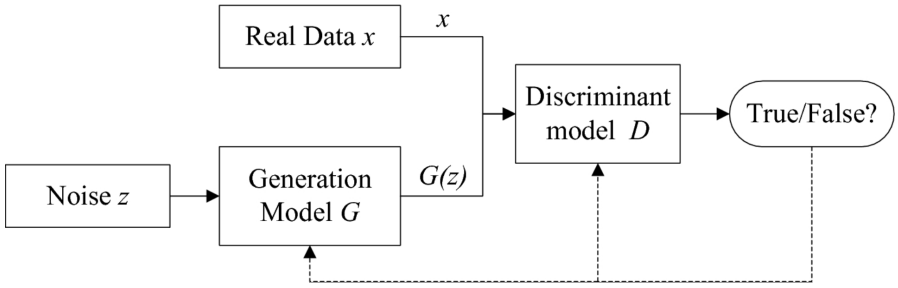


Figure 2. Generative Adversarial Network framework[YLLH18].

Table 7. Overview of semi-supervised learning applications.

Publication	Method	Remarks
	CT	
Chuquicusma et al. [CHBB18]	DCGAN	Lung Nodule
Frid-Adar et al. [FADK+18]	DCGAN/ACGAN	liver lesion
	MR	
Bernudez et al. [BPD+18]	DCGAN	brain
Han et al. [HHR+18]	WGAN	Brain
	X-ray	
Salehinejad et al. [SVD+18]	DCGAN	Chest
Madani et al. [MMKSM18a]	DCGAN	Chest
Madani et al. [MMKSM18b]	DCGAN	Chest
	Retinal fundus imaging	
Gass et al. [GSG12]	DCGAN	-
Lahiri et al. [LAKBM17]	PGGAN	-

generation that follows the same concept of GAN. It takes uniform 100-dimensional noise distribution Z as input and result is a four-dimensional tensor that is used as the beginning of the convolution stack. Thus, Z is projected into a small spatial convolutional representation with several feature maps. For discriminator, last convolutional layer is flattened and fed into sigmoid output. No pre-processing is needed for images, just rescaling. Weights are initialized from a zero-centered normal distribution. For a detailed architecture, please refer to Figure 3.

2.4 Feature extraction techniques

Feature extraction is a crucial preprocessing step as it removes noise and invariant parts from the image, not relevant for a particular classification or regression task. This section reviews some common feature extraction techniques.

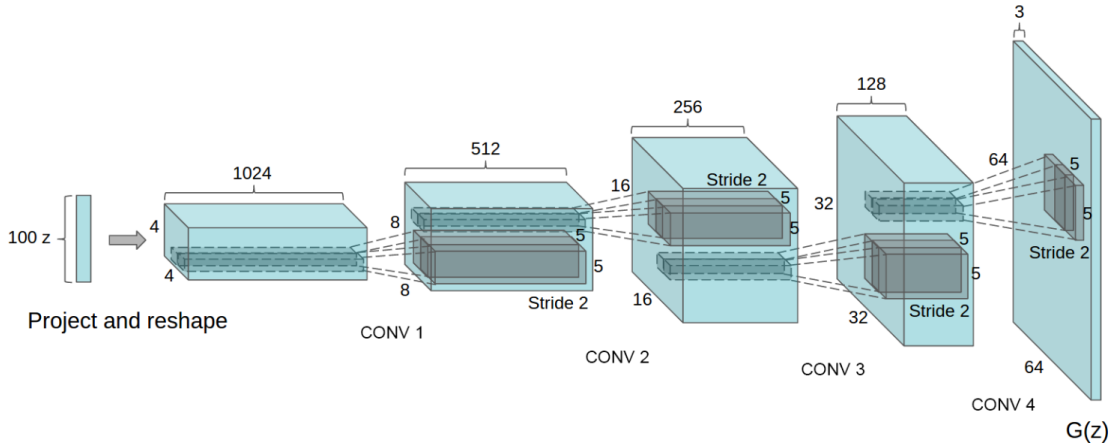


Figure 3. DCGAN architecture [RMC15].

2.4.1 Handcrafted features

Manually designed features are lower dimensional that makes it easier to train a model. However, it requires domain knowledge and a huge amount of time. Considering the past work, many previous studies in medical area extracted shape, color and texture features from the images and trained a shallow model on them [BRF+14] [SWC+11] [IOC+08][RHN+17][RGS+10] but such feature representations cannot be generalized and often have a poor quality. In recent years, it is more preferred to use learned representations of data instead of hand-crafted features.
A natural step is to use the ability of computers to learn the most important features from the input images. It is achieved by models (networks) that consist of many layers and transform data over layers gradually. The deeper they go, the higher level features are obtained. The most successful neural network model while working on images is CNNs. There exists a couple of well-known CNN (VGG [SZ14], ResNet [HZRS16], AlexNet [KSH12]) that are pre-trained on a huge dataset consisting of thousands of colorful images of thousands of classes. Those networks are used in many medical scenarios, as a feature preprocessing step. Specifically: instead of training on the original images, the features from images are extracted from the chosen layer and feed into another classifier. In the end, not only dimensionality is reduced, but more important features are obtained as well. Thus the classification accuracy is usually better than just training on the raw images. This technique is called transfer learning.

2.4.2 Transfer learning

There are two types of transfer learning strategies applied on medical data: 1) use of a pre-trained network 2) fine-tuning a pre-trained network. Antony et al. extracted

features from VGG-16 pre-trained network and after fine-tuning successfully plugged them into linear SVMs to classify knee OA images. Kim et al. [KCRB16] used CNN as a feature extractor and claimed that it outperformed the strategy of fine-tuning. To judge which strategy is better, there are two other papers the readers can refer (Esteva et al. [EKN+17], Gulshan et al. [GPC+16]). The authors fine-tuned a pre-trained VGG network on their dataset and as far as it is known, such results are not achieved by using pre-trained networks as feature extractors.
The pre-trained network followed by SVM as a classifier is a common tool. Anavi et al. [AKG+16] and Liu et al. [LTK16] extracted the features from x-ray images using a fully connected layer of five-layer CNN. Obtained features were then feed in a one-vs-all support vector machine (SVM) for classification. Zhang et al. [ZZM+17], Kumar et al. [KSQ+16], Fonseca et al. [FMW+15] followed the same strategy. The state of the art results were obtained by Liu et al. [LTK16] who used the penultimate fully-connected layer for feature extraction of x-ray images on 193 different classes. Shah et al. [SCNK16] mixed CNN feature descriptors with hashing-forests, specifically: 1000 features, extracted from MRI volumes, were then compressed into descriptors by hashing forests. Burlina et al. [BFJ+16] used the Overfeat pre-trained network [SEZ+13] for feature extraction on the age-related macular degeneration detection task. There are several other works done in nodule detection analysis (Ciompi et al. [CdHvR+15], Van et al. [VGSJC15], Chen et al. [CZP+16]) which also used CNN feature descriptors for their classification tasks.
This work uses VGG-16 network [SZ14] pre-trained on ImageNet [DDS+09] dataset due to it's excellent performance when used as a feature extractor [EKN+17][GPC+16]. The number 16 stands for the number of layers used in the model. The network is known for its simplicity, using only 3×3 dimensional convolutional layers.

2.5 Interpretability

The Following section gives a general overview of some state-of-the-art model interpretability techniques applied in medical imaging.

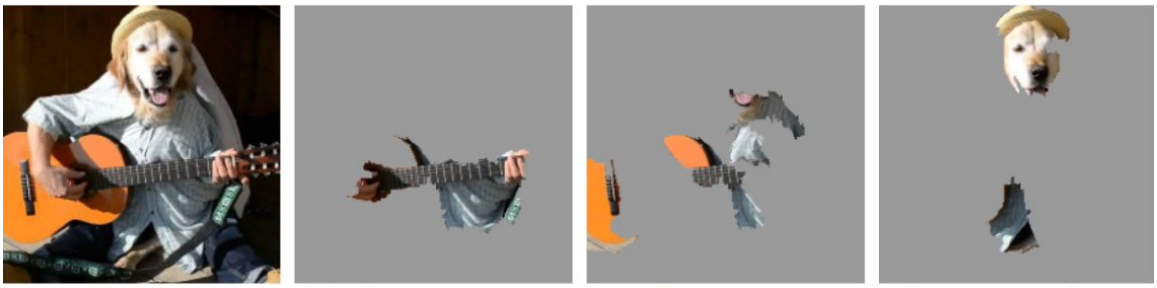
2.5.1 Importance of interpretability

There is no mathematical definition of interpretability. However, in general, interpretability is the degree to which a human understands the cause of a particular decision. The higher the interpretability of a model, the more transparent it is and the easier it is for someone to understand why specific predictions have been made. Model interpretability is crucial, especially in the medical area as an automated diagnosis of the diseases cannot be integrated with real-world scenarios unless there is not enough trust and transparency in the models.

2.5.2 Local model interpretability

There are several methods proposed to address local interpretability problem. In 2016 Ribeiro et al. [RSG16] proposed a technique called LIME that explains predictions of any classifier, including neural networks in an interpretable and straightforward manner. The output of the method is a part of the original image, and it captures the area, that contributed to the particular prediction the most (Figure 4). Specifically, if g is an explanation defined as a model and $\Omega(g)$ is a measure of the complexity of the description of the model f, LIME seeks to minimize locality-aware loss by drawing samples π_x. The explanation is then produced using Equation 1. LIME has also been used to explain model predictions on a mental health classification task [SMB18].

$$h(X) = \underset{(g)\in G}{argmin} L(f, g, \pi_x) + \Omega(g) \tag{1}$$



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4. LIME explaining an image classification prediction made by Google's Inception neural network[RSG16].

Deepminer is proposed by Wu et al. [WZP+18] and it is a framework to identify interpretable explanations for medical predictions. The framework consists of three steps. In the first step, authors trained a basic neural network for breast cancer classification. In the second step, experts were asked to annotate the most class-specific internal units of the trained model. In the final step, annotated units were used to generate explainable predictions by sorting individual unit contributions to each prediction.
Deepbase is another technique proposed by Sellam et al. [SLH+18]. It's a system to inspect neural network behaviors using user-provided hypothesis functions. The main idea in Deepbase is that technique seeks to quantify how the behavior of hidden units is similar to the user-defined functions. Given a group of units and hypothesis, the tool calculates the affinity score between each hidden units and hypothesis pair and statistically measures the difference.
Wu et al. [WPH+18] showed that CNN internal units can learn and detect medical concepts which match the vocabulary used by radiologists. Rajpurkar et al. [RIZ+17]

and Wang et al.[WPL+17] applied class activation maps to explain informative regions related to test set predictions. Hybrid CNN proposed by Zhang et al. [ZXX+17] generated radiological reports if trained on a large image dataset.

2.5.3 Global model Iterpretability

The number of works has been proposed for Global Interpretability. One of them is called Activation Maximization(AM) and is offered by Erhan et al. Idea of AM is the following: It looks for input patterns which maximize the activation of a given hidden unit. The reasoning behind finding those activations is that pattern that arose such activation can be a good representation of what a unit is doing. AM can be interpreted as an optimization problem. If θ is a set of neural network parameters, and $h_{ij}(\theta, x)$ denotes activation of unit i from layer j then, given input x, what AM looks for is depicted in the Equation 2.

$$x' = arg \max_{xs.t||x||=p} h_{i,j}(\theta, x) \tag{2}$$

AM has been used as a global interpretability tool for breast cancer prediction by Graziani et al. [GAM18b].
Pereira et al. [PMM+18] proposed a technique that can interpret model predictions globally in brain tumor segmentation and penumbra estimation task. The main focus of that work is on the interpretation of automatically learned features- basically, authors study the interpretation of features from Magnetic Resonance Imaging sequences (MRI). To do this, authors couple RBM representation model with RF and jointly considered correlations between images, features and target variables.

3 Methodology

The following section gives a very detailed overview of the performed experiments and the proposed framework. It starts by explaining the dataset and the baseline of the work, followed by presenting feature extraction and data augmentation methods used in different data labeling techniques. The chapter concludes by giving a thorough overview of the proposed CNN architecture and model interpretability methods.

3.1 Proposed framework

The proposed study consists of two stages. In the first stage, a CNN network is proposed for breast cancer classification, which outperforms other papers working on the same dataset for the same problem [CRBG$^+$14][JM16]. To make the comparisons valid, no data augmentation or feature extraction techniques are applied on the original images, and the same dataset with the same training and testing instances are feed in the proposed CNN network.
The second stage, which is the main contribution of the study, refers to the case when there is no way of having a broad set of labeled instances. In contrast to the first stage, it is assumed that only a small portion of data is labeled while the more extensive set remains unlabeled. The proposed framework labels the rest of the instances and trains CNN with the newly labeled set. To simulate having a lack of data, the truth files for most of the instances are covered. The ratio between labeled and unlabeled examples is 1:10. In the first path, the aim is to enlarge the size of the labeled set. For this reason, DCGAN is adopted for new image synthesis. Generated images are appended in the training set, and VGG-16 pre-trained network is used to extract features from both training and testing instances. In the second path, 10 Gradient Boosting (GB) models are used to train the rest of the unlabeled instances. After data labeling, CNN is trained on the mixed set of both, labeled and newly labeled instances, and performance is tested on the same test set, used in the first stage. In the third and the final path, model interpretability technique is applied to each of the testing instances in order to understand why specific predictions are made. The overall framework is shown in Figure 5.

3.2 Dataset

The original annotated dataset obtained by Cruz et al. [CRBG$^+$14] consists of 162 WSI images that come from patients diagnosed with IDC. Initially, the slides were randomly split into three different subsets: 84 training cases and 29 validation cases for parameter exploration, and 49 test cases for final evaluation. Each WSI was divided into non-overlapping image patches of size 50x50 via grid sampling and ones with fatty tissue, and slide background was removed. Image patch was labeled as positive if it belonged

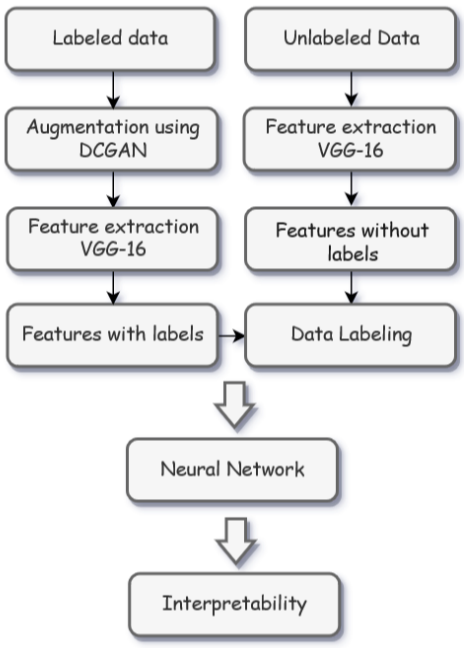


Figure 5. The Overall framework for the proposed study

to more than 80 % of the annotated mask, otherwise labeled as negative. The sampling procedure is illustrated in Figure 6.
The final patch-based dataset consists of 82,883 (A_1) and 31,352 (A_2) instances for training and validation, and 50,963 (A_3) instances for testing. Dataset is highly unbalanced with around 70% of negative instances. This work experimentation with the patch-based dataset and train-test-validation sets is identical to the one in the baseline paper [CRBG$^+$14]. The example of the sampled patches is shown in Figure 7. To simulate the lack of unlabeled data, 12,000 (10% of A_1 and A_2) instances were randomly taken from A_1 and the rest of the instances in the training and validation sets are assumed to be unlabeled. Beyond this point, $labeled\ data$ is referred to that 10% of instances instead of the originally labeled set unless specified. Images were downscaled to the dimensions 32x32 to make them compatible with the proposed methods.

3.3 Baseline

Baseline for this work applies a 3-layer CNN to detect malignant IDS tissue regions. The network receives sampled patches as input and it outputs a probability for a cell being malignant. Network output (probabilities) is then stitched onto the original WSI to obtain the IDC probability map over WSI. The framework and the architecture used in

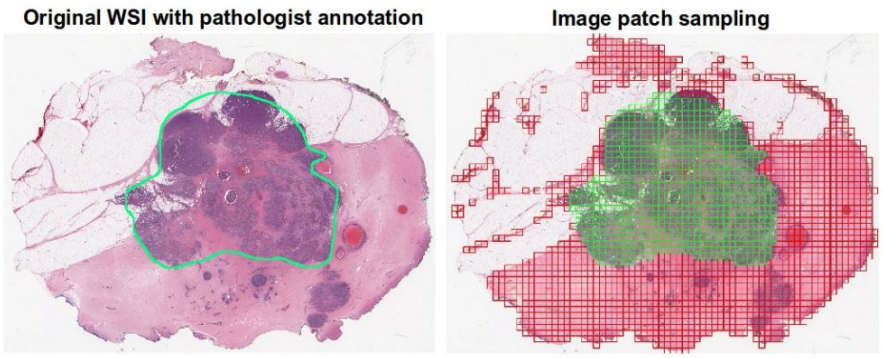


Figure 6. Image patch sampling process. The original WSI with annotations is shown to the left and splits of image patches to the right. The white parts are removed. Red color corresponds to the positive patch of IDC, while green is a negative example [CRBG+14].

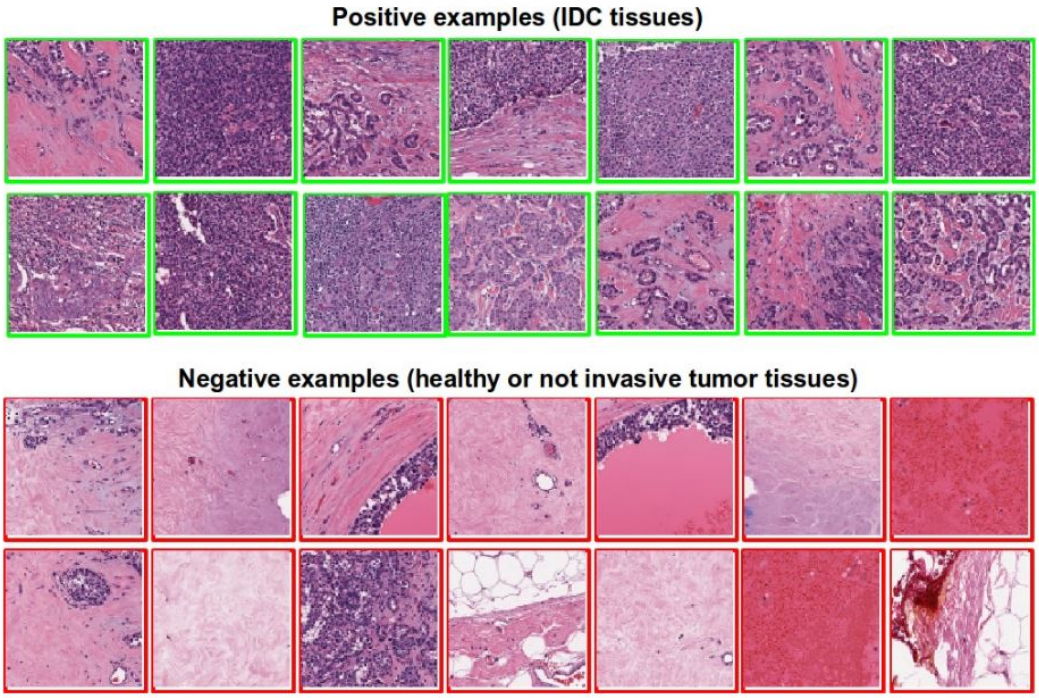


Figure 7. Examples of positive and negative patches in the dataset [CRBG+14].

the paper are shown in Figure 8.

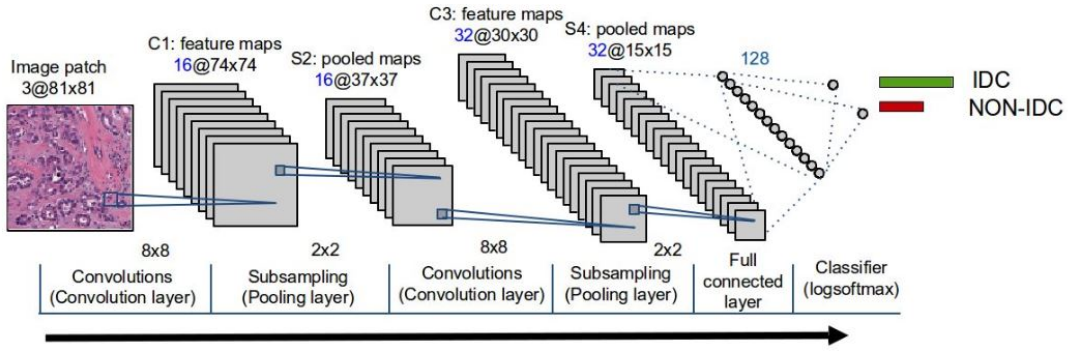


Figure 8. Overall framework of the baseline study [CRBG+14].

Another experiment pursued by Janowczyk et al. [JM16] obtained better results than the original paper by using a well-known AlexNet architecture. The Study showed that relatively smaller and compact network could produce comparable results.

3.4 Data augmentation

DCGAN (Section 2.3.2) is adopted for data augmentation due to its good training stability. DCGAN is trained on each class separately to make the generated images more useful. Obtained synthetic images are appended in the labeled set.

3.5 Feature extraction

Synthetic and original images are converted to YUV color space. YUV is a color encoding system that takes human perception into account. By enabling transmission errors, images are masked better than just a "direct" RGB-representation. VGG-16 pre-trained network on the ImageNet dataset is proposed as a feature extraction tool. The GlobalAveragePooling is applied to the four internal convolutional layers and the result is concatenated into one vector [RSIK18]. After this step, the high-level feature set is returned with a significantly reduced dimension. Refer to Figure 9 for more details.

3.6 Data labeling

The labeled set is further split into training (B_1), validation (B_2) and testing (B_3) sets and experiments are evaluated on the validation instances. Several classifiers are applied to find the best performing model to label the rest of the instances (A_1, A_2), and it will become the primary model for data labeling in this study.
LightGBM, the ML technique for either regression or classification problems, is a

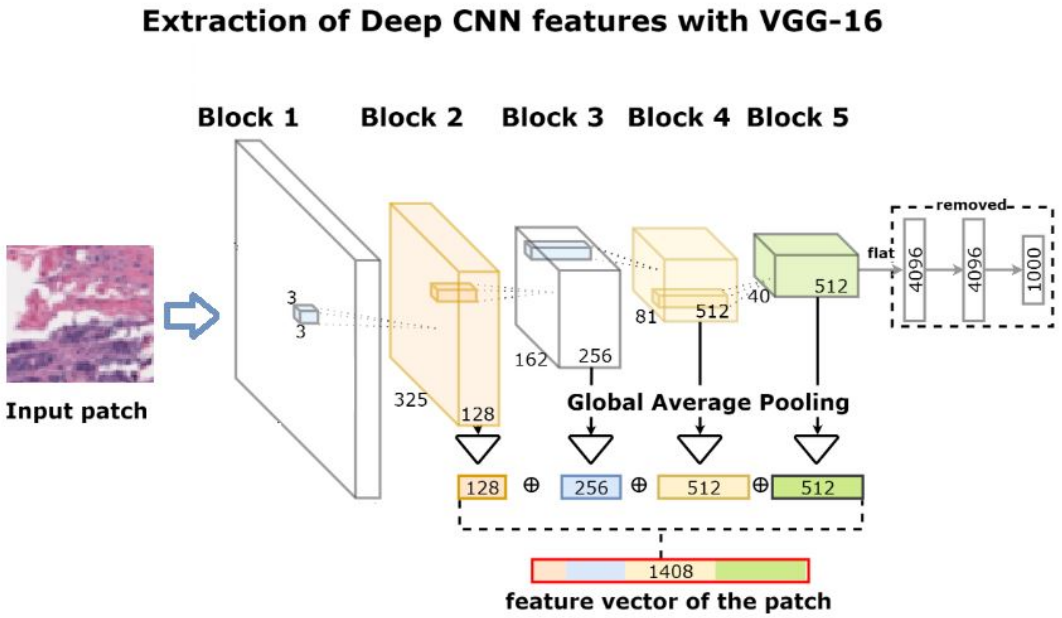


Figure 9. Overview of the network architecture for deep feature extraction [RSIK18].

framework of GB models. Predictions are produced in the form of weak classifiers, that are presented as decision trees. The loss is optimized on the validation dataset using the log loss depicted in Equation 3.

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i log p_i + (1-y_i) log(1-p_i)] \tag{3}$$

N is the number of samples; y_i is a binary indicator of whether or not label j is the correct classification; and p_i is the model probability of assigning label j to the example i. After all the instances are labeled, they are appended in the labeled set and used to train the CNN.

3.7 Neural network

For the best architecture and parameter optimization, NNI is adopted. A detailed explanation of the platform is provided in AutoML Section 2.2. First, NNI is applied to the initially labeled dataset (A_1, A_2) to find the best architectural design and make comparable results concerning the baselines [CRBG$^+$14][JM16]. To address the main contribution that assumes the case when a small portion of the labeled data is available, newly labeled instances, along with the initially labeled set are feed into the NNI platform

again. The best architecture found in both cases consists of three convolutional layers and two fully connected layers, applying ReLU activation and batch normalization before each convolutional layer, following max-pooling technique after each layer of CNN. The number of kernels for each layer was set to 64. The detailed architecture is shown in Figure 10.

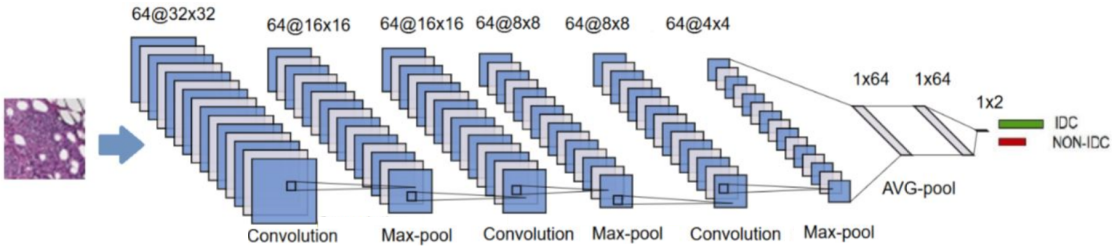


Figure 10. A visualization of the CNN network for the mixed data

3.8 Model interpretability

Developing trust in individual model predictions is an important problem, especially for decision making in the medical field. Moreover, it is also crucial to evaluate a model as a whole, before integrating it and applying into real-world problems. Understanding the logic behind a network will lead to more transparency, better design, and quicker experiments. To gain more confidence, this work adopts neural network inspection techniques. Such interpretable systems will help researchers to identify failure models, which allows them to focus their effort into weak side development.
To understand how the CNN is learning the newly labeled set, two interpretability techniques are adopted: Grad-Cam proposed by Selvaraju et al. [SCD+17] and Regression Concept Vectors proposed by Graziani et al. [GAM18a]. The following section gives an overview of the use of those techniques in this study.

3.8.1 Grad-Cam

Grad-Cam is a localization map that generates local visual explanations on the input image and identifies regions, which were the most important while taking a particular decision. Grad-Cam uses gradient information flowing in the last convolutional layer, as it is assumed that information is lost in fully-connected layers. To obtain the localization map, the gradient of the score y^c is computed for the class c concerning the feature map A^k. These gradients flowing back are then global-average-pooled, and neuron importance

weights are computed (Equation 4).

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\nabla y^c}{\nabla A_{ij}^k} \tag{4}$$

After performing a weighted combination of forward activation maps, followed by ReLU activation, localization heat-map is obtained. (Equation 5).

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \tag{5}$$

Localization maps acquired from proposed CNN is shown in Figure 11. The first and the third rows are negative and positive class examples respectively and the second and fourth rows are the corresponding heat-maps. Both tumor and non-tumor heat-maps include nuclei-resembling shapes, but Grad-Cams of positive examples seem to capture more nuclei parts. However, more research is needed to interpret the discriminant factors between those classes.

3.8.2 Regression Concept Vectors

Saliency methods like Grad-Cam are one of the most popular explanation techniques, but they have a couple of limitations: 1) since a saliency map is conditioned on only one image, humans have to manually assess each picture to draw a class-wide conclusion, 2) users do not have control over what concepts are illustrated in the heat-maps. Those limitations are also depicted in Figure 11. Firstly, it is not clear exactly how the saliency maps differ across the classes. Secondly, heat-maps are not very informative to identify if there is any concept that is a reason for particular class prediction. Moreover, it is difficult to draw a broader conclusion unless someone is not a doctor. This study uses directional derivatives of Regression Concept Vectors to measure the influence of user-defined concepts in the network output. The following section explains how the textual concepts are obtained and how they are used for explaining network predictions.

3.8.2.1 Framework

Texture analysis is a useful way of extracting relevant information from images and an essential step in discriminating between normal and benign tissues that show early signs of breast cancer [KASS14]. The texture itself is interpreted as the distribution of gray levels in the images. Haralick et al. propose one of the most famous approaches to texture analysis [HS$^+$73] and it is based on the Gray-level Co-occurrence Matrices (GLCM), obtained from the images. GLCM is computed for various angular moments and distances and in total, thirteen textual features can be extracted. Each cell (i, j)

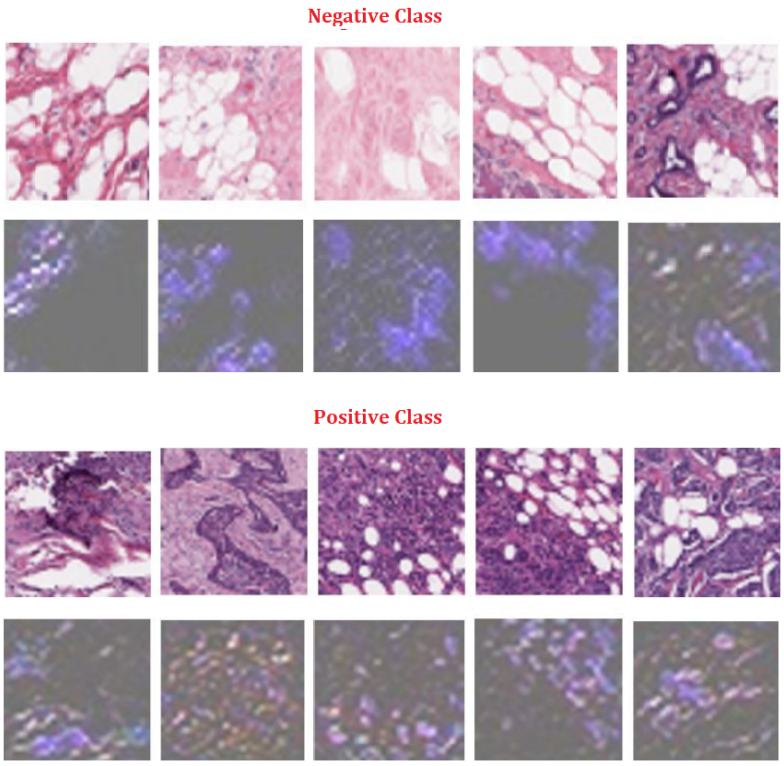


Figure 11. Grad-Cam for testing patches

in GLCM corresponds to the number of occurrences in the gray levels of i, j. In this study, three out of thirteen concepts are calculated, and they refer to the Nottingham Histologic Grading System (NHG) [EE02]. Only the following concepts are considered: Contrast (f_1) – a measure of gray-tone linear-dependencies in an image, Correlation (f_2) – a number of local variations presented in an image and ASM (f_3) – a measure of homogeneity of the image. In a homogeneous image, there are a very few dominant gray-tone transitions.

$$f_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)^2 \tag{6}$$

$$f_2 = \sum_{k=0}^{N_g-1} k^2 p_{x-y}(k) \tag{7}$$

$$f_3 = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i,j)p(i,j) - \mu_x\mu_y}{\nabla_x\nabla_y} \tag{8}$$

Concepts are calculated from nuclear segmentation dataset [KVS+17], for which labels of tumorous and non-tumorous regions are not given. The dataset contains different organs, but 100 image patches are extracted from WSI breast tissue only. Features $\Phi_l(x_j)$, from layer l of the trained CNN, are extracted from the same set of images and linear regression (LR) is solved for each concept c_j separately. LR is fitted so that it follows increasing values of the concept measurements. After this stage, three directional vectors are obtained those are called Regression Concept Vectors (RCVs). RCVs are used for sensitivity score calculation, and for bidirectional relevance score measurement, that is used for global interpretability of the proposed CNN. General steps of this process are depicted in Algorithm 1 and the framework can be found in Figure 12.

Algorithm 1 Regression Vector Concepts

x': nuclei segmented images
Φ: trained network
x: original patches
procedure INTERPRETABLESCORES
    for each image in x' **do** calculate concept c_j
    end for
    for each concept in c **do** R, $\vec{v_c^l}$ =LR$(\Phi_l(x'_j), c_j)$
    end for
    for each testing patch in X **do** $\mathrm{S}^l_{C,i} = \nabla f(\Phi^l(x_i) \times \vec{v_c}$
    end for
    Br=$R^2 \times (\frac{\mu}{\sigma})$
    return S_c, Br
end procedure

3.8.2.2 Sensitivity scores

Sensitivity Scores $S^l_{c,i}$, $c = 1, 2, 3$ are computed for each testing input (x_i, y_i) separately (Equation 9).

$$S^l_{C,i} = \nabla f(\Phi^l(x_i) \times \vec{v_c} \tag{9}$$

Each score represents the network sensitivity along the direction ($\vec{v_c}$) of increasing values of the concept. For instance, if the concept measure in the input increases, the probability

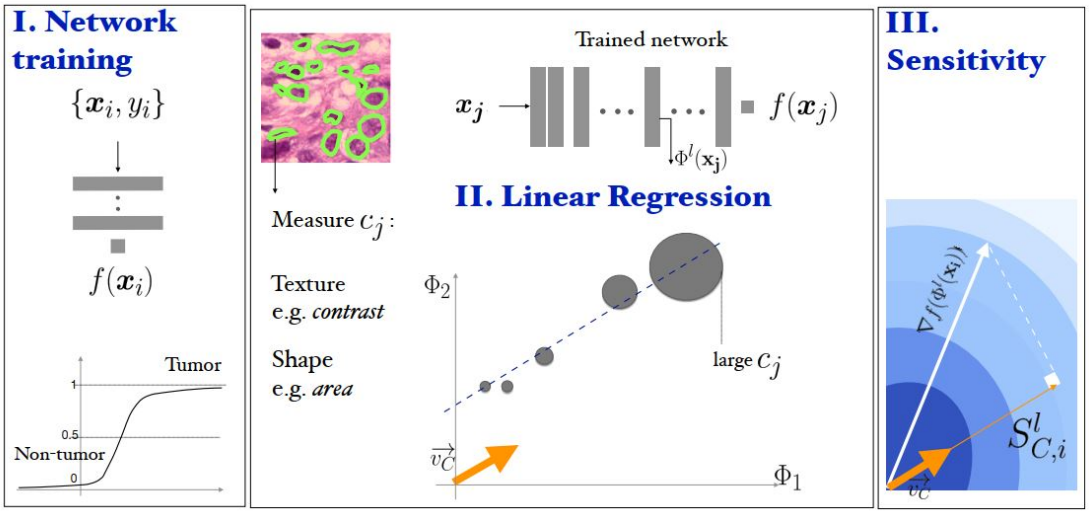


Figure 12. Regression Concept Vectors for Bidirectional Explanations in Histopathology. I. The CNN is Trained on the newly labeled data. II. Linear Regression is solved on $\Phi_l(x_j)$ at layer l for each concept c_j. III. Sensitivity scores are computed for each testing instance x_i as the derivative of the network prediction $f(x_i)$ on the direction of the RCV [GAM18a].

of being malignant cell might decrease, increase or remain unchanged. The magnitude of the score represents the rate of change, whereas the sign shows the direction of change. Sensitivity scores can be computed for all the testing instances in the test set.

3.8.2.3 Bidirectional Relevance Score

Bidirectional Relevance (BR) score is used for global interpretability of the model that is a ratio between coefficient of determination of the least squares regression R and a standard deviation of the Sensitivity Scores (Equation 10).

$$Br = R^2 \times (\frac{\mu}{\sigma}) \tag{10}$$

A negative value of Br score of the concept i indicates that i negatively contributed to the network predictions. Thus, it is the reason of non-tumor cases, whereas positive Brs suggest the concepts that arise positive(tumor) predictions.

3.9 Experimental setup

Experiments are done using the machine with Intel(R) Core(TM) i5-7300U CPU processor. DCGAN was trained using Stochastic Gradient Descent as an optimizer for 1200 epochs with the batch size of 16 and learning rate equal to 0.0003. Loss function applied

is a binary cross-entropy. Parameters are the same for both discriminator and generator. Training took ten days for both classes.
Feature extraction from VGG-16 pre-trained network took two days. For Gradient Boosting models, parameters are estimated using a grid search. Found learning rate, number of leaves, maximum depth, feature fraction, bagging fraction values are 0.1, 191, 7, 0.46, 0.69, respectively. For the best neural architecture search, NNI was trained using AUC as a loss function and Stochastic Gradient Descent as an optimizer. Best hyperparameter values in both networks are the same: learning rate, epoch, batch size, momentum, decay-0.01, 80, 32, 0.9, 1e-5, respectively. The model was trained using early stopping criteria with patience set to ten. For the best neural architecture search and hyperparameter tunning, budget assigned was set to 24 hours.

4 Results

The following section explains the experimental design, performance measures used to evaluate classification methods, the effects of data labeling, proposed CNN architecture, and the model interpretability technique.

4.1 Experimental design

Image patches are classified as one or zero. In the case of training the CNN on the originally labeled patches on the full dataset, the same training (A_1), validation (A_2) and test (A_3) sets are used for parameter exploration and for testing the model performance as it was used in the baseline. In the case of the semi-supervised learning, however, a small subset is constructed accounting for the 10% of the training (A_1) and validation (A_2) instances in total. Moreover, this 10% is further subsetted to construct training (B_1), validation (B_2) and testing folders (B_3) with the proportion of 8:2:90 to simulate labeling the more significant portion of instances. The Best labeling model is then applied to the rest of the A_1 and A_2 instances. The newly labeled set (A_1, A_2) combined with the initial labeled data is used to train CNN and performance is tested on the A_3 testing set.

4.2 Performance measures

Precision (Pr) is a proportion of IDC detected from the total number of IDC areas. Recall (Rc) or Sensitivity (Sen) is a proportion of IDC correctly predicted out of the total number of predicted IDC. Specificity (Spc) is a proportion of Non-IDC regions correctly predicted out of the total number of Non-IDC parts. Because the dataset is highly unbalanced, the primary labeling measure is the balanced accuracy, which is the same as the area under the curve (Equation 11) for the binary labels. Other measures used in the evaluation are the F-measure (Equation 12) and the accuracy, where the accuracy is the percentage of the examples correctly classified.

$$BAC = \frac{Sen + Spc}{2} \tag{11}$$

$$F1 = \frac{2 \times PR \times Rc}{Pr + Rc} \tag{12}$$

4.3 Data Labeling method

Features are extracted from the VGG-16 pre-trained network, and several classifiers are evaluated to find the best performing data labeling technique. Table 8 shows the AUC and the accuracy of the newly labeled dataset using different machine learning models.

As it is seen from Table 8, GB and RF classifiers give the best results, and between those, GB model achieves a better performance. Table 9 shows a different number of GB as an ensemble model. The results are almost the same, but 10 GB is slightly better than the others.

Table 8. Labeling performance across models.

Model	AUC	ACC
RF	0.8321	82.32
GB	0.8373	84.86
DT	0.7558	75.43
AdaBoost	0.8308	81.91
Gaussian	0.7581	71.06

Table 9. Number of GB vs. performance

N	AUC	ACC
4	0.8381	82.70
6	0.8386	82.83
8	0.8385	82.81
10	0.8387	82.9
12	0.8386	82.86

Table 10 and 11 show data labeling performance of 10 GB models using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Since LDA only has at most 2 classes, only one projection is produced that explains why the same results are obtained across different number of components. The performance is deficient in comparison to PCA as well. In overall, It can be concluded that VGG-16 extracts higher level features than PCA as there is more than a percent difference in data labeling performance by 10 GB models.

Figure 13 shows the AUC of 10 GB models when a different percentage of synthetic images are added in the labeled set. Features are extracted from VGG-16 pre-trained network. For instance, 30 % (of 12000) means 3600 instances, 50 % means 6000 instances, and so on. The highest AUC, 0.8474, is achieved when 100% of generated images are added. In this case, the training size is doubled. It is the best result achieved in data labeling, and the CNN is trained on this newly labeled set alongside the initially labeled data. Examples of the generated synthetic images are shown in Figure 14.

4.4 CNN model performance

Figure 15 shows CNN performance using only different amount of labeled data. The results are also compared with a various number of labeled and fix size of newly labeled

Table 10. PCA for feature extraction.

Components	AUC	ACC
50	0.8269	83.87
100	0.8272	83.96
200	0.8247	83.78
250	0.8241	83.58

Table 11. LDA for feature extraction

K	ACC	AUC
50	0.4651	0.6412
100	0.4651	0.6412
200	0.4651	0.6412
250	0.4651	0.6412

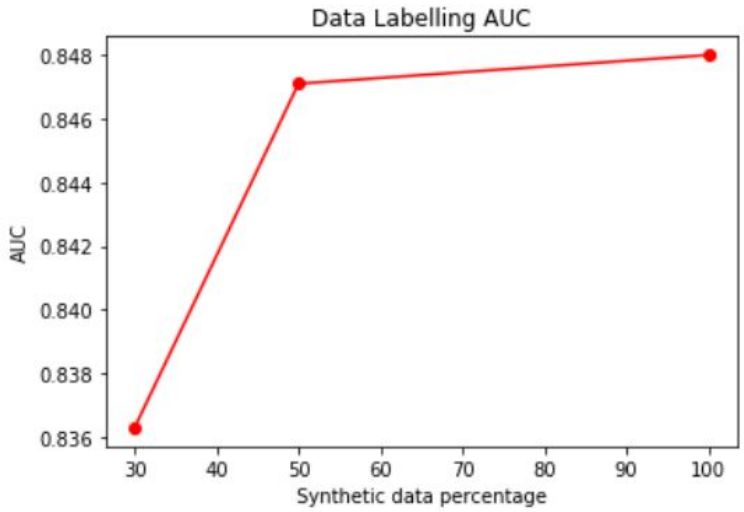


Figure 13. Data Labeling performance of 10 GB models with different number of synthetic images added in the training set.

data. The performance gradually increases as the number of initially labeled instances grows. It is also clearly seen that adding newly labeled examples to the labeled set increases the performance and hits the highest score – 0.7729 F-measure, 0.8649 AUC with 12 000 (10%) instances labeled and the remaining 90% newly labeled.

Table 13. Performance on originally labeled data

Method	F-measure	AUC
Cruz et al.	0.718	0.8423
Janowczyk et al.	0.7648	0.8468
Our approach	0.7923	0.8696

To compare the performances using all the originally labeled data with the baselines, the truth files are uncovered to train the CNN. The F-measure of originally labeled data is 0.7923, whereas Cruz et al. [CRBG$^+$14] reported 0.718 and the best achieved by [JM16] is 0.7648. The results are shown in Table 13. Further, there is a 2% difference in the F-measure when all data is labeled VS. newly labeled, that is to be expected as originally labeled set contains more accurate information.

4.5 Model interpretability

As it was shown in Figure 11, Nucleipleomorphism seems to be the main factor affecting the output prediction. Heatmaps of tumor patches capture nuclei variations that focus on

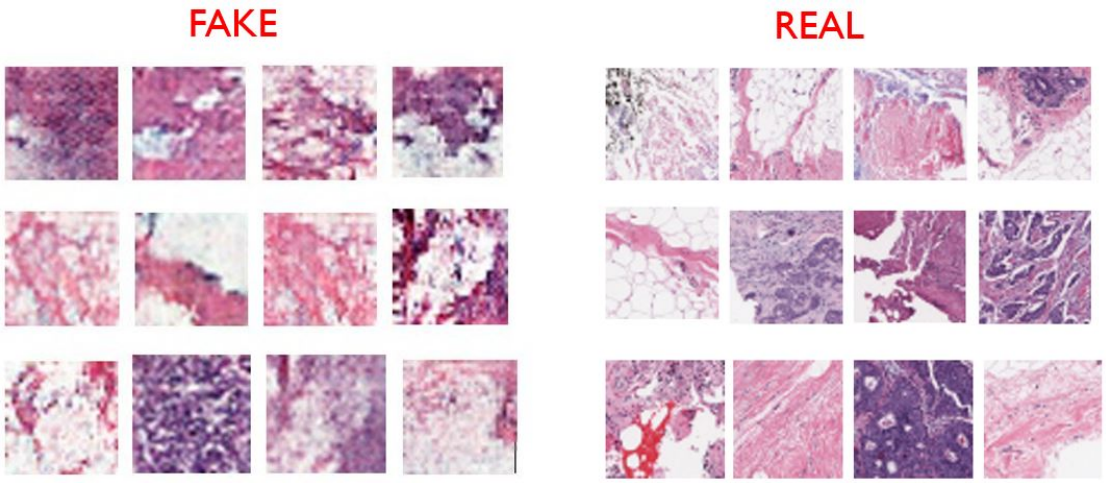


Figure 14. Synthetic images generated by DCGAN

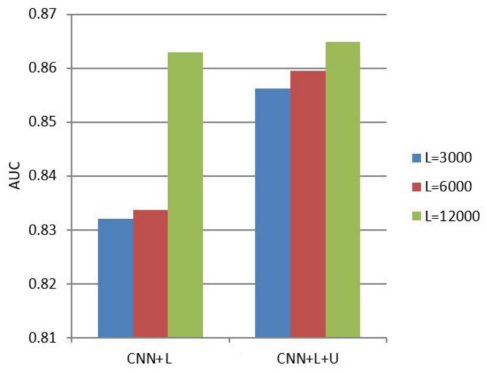


Figure 15. AUC using labeled data only and mixed data

nuclei with various shapes and sizes. Figure 16 shows the obtained Br concept measures. Based on the results, Contrast is a concept relevant for classification, and this is also by the NHG grading system, that identifies hyperchromatism as a sign of nuclear atypia. Moreover, Br also shows that concept Correlation negatively affects classification results and causes negative predictions. Thus, when this concept in nuclei sufficiently increases, the prediction of the input patch switches from class IDC to class Non-IDC. Sensitivity scores mirror Br score results that strengthens the conclusions.

4.6 Discussion

In the following study, a semi-supervised, CNN network for breast cancer prediction is developed, that can use a large amount of unlabeled set to improve the performance. In

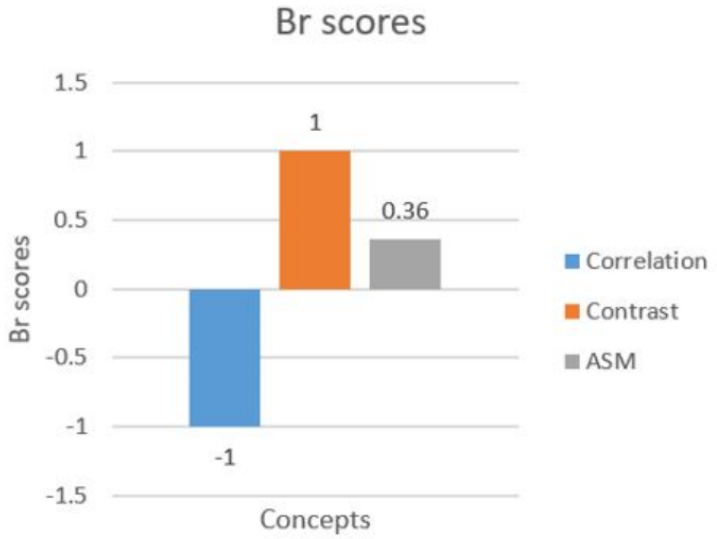


Figure 16. Br concept measures

comparison to other image classification cases, labeled medical images are complicated to obtain. The provided method is a step closer towards deep learning integration with medical diagnostics. The approach introduced in this study makes use of a tiny portion of labeled instances, and in the scenarios where not enough labeled data is available, existing sources of the unlabeled set can be utilized.

To make use of unlabeled data and automatically label it as accurately as possible, different settings are tested. The best result achieved is using 10 GB models, when 100% of synthetic images are added in the initial labeled set. Features extracted from VGG-16 gave the best results in comparison to other dimensionality reduction techniques, and the best $labeling$ performance achieved in this study is 0.8474 AUC. From Figure 14, it can be concluded that the amount of labeled data has a significant influence on the performance of CNN. Both the AUC and the F-measure increase gradually as the size of the labeled set increases. Based on those experiments it can also be concluded that unlabeled data provides some extra information useful for classification, as training CNN using mixed data resulted in a better performance, rather than training on an initially labeled set only. The highest CNN performance achieved on the newly labeled data is the F-measure – 0.7729 and the AUC – 0.8649. The same CNN was trained with the originally labeled dataset and compared to two other baselines [CRBG$^+$14] [JM16] and it was shown that provided CNN architecture gives a considerable performance boost: 0.7923 F-measure, 0.8696 AUC. There is a 2% difference in the F-measure when all data is labeled vs. newly labeled, as initial labels contain more accurate information for the network. The proposed CNN was also inspected by two techniques: Grad-Cam and

Regression Concept Vectors. The techniques successfully marked regions, relevant for classification, and also identified two textual concepts – Correlation and Contrast which were found to be the reason for positive and negative class predictions.

5 Conclusion

This study developed a framework that enables users to use a small amount of labeled data and a much larger set of unlabeled data to train a CNN. This was done by augmenting the dataset using GANs and feature extraction with VGG-16 pre-trained network. The unlabeled set was labeled with proposed ten gradient boosting models. The results showed that unlabeled data could increase the performance (Figure 15). The CNN performance was also compared on mixed (labeled+unlabeled) data and fully labeled data, and as it was expected, there was a 2% performance increase in the F-measure when all data was labeled. The unlabeled set will never be able to replace labeled data as it contains more accurate information, but there are some situations when obtaining the completely labeled set is not possible, especially in the medical domain and this work pertains to such cases. It enables users to use an unlabeled set and train the CNN, while still achieving acceptable performance – 0.8649, 0.7729 for the AUC and the F-measure, respectively. Further, the study adopted two interpretability techniques, that can adequately inspect the proposed CNN network both locally and globally and identify concepts, that positively and negatively affect classification results. Future work might include increasing labeling performance and advancing neural network inspection methods.

References

[AAL+15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[AAL+16] Le An, Ehsan Adeli, Mingxia Liu, Jun Zhang, and Dinggang Shen. Semi-supervised hierarchical multimodal feature and sample selection for alzheimer's disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 79–87. Springer, 2016.

[ABP16] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016.

[AKCK17] Varghese Alex, Mohammed Safwan KP, Sai Saketh Chennamsetty, and Ganapathy Krishnamurthi. Generative adversarial networks for brain lesion detection. In *Medical Imaging 2017: Image Processing*, volume 10133, page 101330G. International Society for Optics and Photonics, 2017.

[AKG+16] Yaron Anavi, Ilya Kogan, Elad Gelbart, Ofer Geva, and Hayit Greenspan. Visualizing and enhancing a deep learning framework using patients age and gender for chest x-ray image retrieval. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, page 978510. International Society for Optics and Photonics, 2016.

[BDP+11] Kayhan N Batmanghelich, H Ye Dong, Kilian M Pohl, Ben Taskar, Christos Davatzikos, et al. Disease classification and prediction via semi-supervised dimensionality reduction. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1086–1090. IEEE, 2011.

[BDW+18] Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, and Hayit Greenspan. Chest pathology identification using deep feature selection with non-medical training. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):259–263, 2018.

[BFJ+16] Philippe Burlina, David E Freund, Neil Joshi, Y Wolfson, and Neil M Bressler. Detection of age-related macular degeneration via deep learn-

ing. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 184–188. IEEE, 2016.

[BKZ+18] Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc Le. Understanding and simplifying one-shot architecture search. In *International Conference on Machine Learning*, pages 549–558, 2018.

[BPD+18] Camilo Bermudez, Andrew J Plassard, Larry T Davis, Allen T Newton, Susan M Resnick, and Bennett A Landman. Learning implicit brain mri manifolds with deep learning. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105741L. International Society for Optics and Photonics, 2018.

[BRF+14] Catarina Barata, Margarida Ruela, Mariana Francisco, Teresa Mendonça, and Jorge S Marques. Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Systems Journal*, 8(3):965–979, 2014.

[BTI+13] Tom Brosch, Roger Tam, Alzheimer's Disease Neuroimaging Initiative, et al. Manifold learning of brain mris by deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 633–640. Springer, 2013.

[CB16] Christoforos C Charalambous and Anil A Bharath. A data augmentation methodology for training machine/deep learning gait recognition algorithms. *arXiv preprint arXiv:1610.07570*, 2016.

[CBC+] Anca Ciurte, Xavier Bresson, Olivier Cuisenaire, Nawal Houhou, Sergiu Nedevschi, Jean-Philippe Thiran, and Meritxell Bach Cuadra. Semi-supervised segmentation of ultrasound images based on patch representation and continuous min cut. *PloS one*, 9.

[CdBP19] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 2019.

[CdHvR+15] Francesco Ciompi, Bartjan de Hoop, Sarah J van Riel, Kaman Chung, Ernst Th Scholten, Matthijs Oudkerk, Pim A de Jong, Mathias Prokop, and Bram van Ginneken. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box. *Medical image analysis*, 26(1):195–202, 2015.

[CHBB18] Maria JM Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 240–244. IEEE, 2018.

[CRBG+14] Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *Medical Imaging 2014: Digital Pathology*, volume 9041, page 904103. International Society for Optics and Photonics, 2014.

[CZH18] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.

[CZP+16] Hao Chen, Yefeng Zheng, Jin-Hyeong Park, Pheng-Ann Heng, and S Kevin Zhou. Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 487–495. Springer, 2016.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[DRK+14] Eva Dittrich, Tammy Riklin Raviv, Gregor Kasprian, René Donner, Peter C Brugger, Daniela Prayer, and Georg Langs. A spatio-temporal latent atlas for semi-supervised learning of fetal brain segmentations and morphological age estimation. *Medical image analysis*, 18(1):9–21, 2014.

[DSBJ11] Carol DeSantis, Rebecca Siegel, Priti Bandi, and Ahmedin Jemal. Breast cancer statistics, 2011. *CA: a cancer journal for clinicians*, 61(6):408–418, 2011.

[EE02] Christopher W Elston and Ian O Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. cw elston & io ellis. histopathology 1991; 19; 403–410: Author commentary. *Histopathology*, 41(3a):151–151, 2002.

[EKN+17] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[FADK+18] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[FDI+11] Roman Filipovych, Christos Davatzikos, Alzheimer's Disease Neuroimaging Initiative, et al. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (mci). *NeuroImage*, 55(3):1109–1119, 2011.

[FMW+15] Pablo Fonseca, Julio Mendoza, Jacques Wainer, Jose Ferrer, Joseph Pinto, Jorge Guerrero, and Benjamin Castaneda. Automatic breast density classification using a convolutional neural network architecture search procedure. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 941428. International Society for Optics and Photonics, 2015.

[GAA+17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[GAM18a] Mara Graziani, Vincent Andrearczyk, and Henning Müller. Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 124–132. Springer, 2018.

[GAM18b] Mara Graziani, Vincent Andrearczyk, and Henning Müller. Visual interpretability for patch-based classification of breast cancer histopathology images. 2018.

[GDDM14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[GPC+16] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi

Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[GSG12] Tobias Gass, Gábor Székely, and Orcun Goksel. Semi-supervised segmentation using multiple segmentation hypotheses from a single atlas. In *International MICCAI Workshop on Medical Computer Vision*, pages 29–37. Springer, 2012.

[GZB+17] Lin Gu, Yinqiang Zheng, Ryoma Bise, Imari Sato, Nobuaki Imanishi, and Sadakazu Aiso. Semi-supervised learning for biomedical image segmentation via forest oriented super pixels (voxels). In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 702–710. Springer, 2017.

[HAGEB16] Ehsan Hosseini-Asl, Georgy Gimel'farb, and Ayman El-Baz. Alzheimer's disease diagnostics by a deeply supervised adaptable 3d convolutional network. *arXiv preprint arXiv:1607.00556*, 2016.

[HB04] Ju Han and Bir Bhanu. Statistical feature fusion for gait-based human recognition. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

[HGYR17] Zeshan Hussain, Francisco Gimenez, Darvin Yi, and Daniel Rubin. Differential data augmentation techniques for medical imaging classification tasks. In *AMIA Annual Symposium Proceedings*, volume 2017, page 979. American Medical Informatics Association, 2017.

[HHR+18] Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. Gan-based synthetic brain mr image generation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 734–738. IEEE, 2018.

[HK16] Sangheum Hwang and Hyo-Eun Kim. Self-transfer learning for weakly supervised lesion localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 239–246. Springer, 2016.

[HLG16] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, 2016.

[HS+73] Robert M Haralick, Karthikeyan Shanmugam, et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[HTC+17] Bo Hu, Ye Tang, Eric I Chang, Yubo Fan, Maode Lai, Yan Xu, et al. Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks. *arXiv preprint arXiv:1711.11317*, 2017.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[ILTT10] Juan Eugenio Iglesias, Cheng-Yi Liu, Paul Thompson, and Zhuowen Tu. Agreement-based semi-supervised learning for skull stripping. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 147–154. Springer, 2010.

[IOC+08] Hitoshi Iyatomi, Hiroshi Oka, M Emre Celebi, Masahiro Hashimoto, Masafumi Hagiwara, Masaru Tanaka, and Koichi Ogawa. An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm. *Computerized Medical Imaging and Graphics*, 32(7):566–579, 2008.

[JDG12] Andrew R Jamieson, Karen Drukker, and Maryellen L Giger. Breast image feature learning with adaptive deconvolutional networks. In *Medical Imaging 2012: Computer-Aided Diagnosis*, volume 8315, page 831506. International Society for Optics and Photonics, 2012.

[JM16] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

[KALL17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[KASS14] B Kishore, R Vijaya Arjunan, Rupsa Saha, and Siva Selvan. Using haralick features for the distance measure classification of digital mammograms. *International Journal of Computer Applications*, 6(1), 2014.

[KCRB16] Edward Kim, Miguel Corte-Real, and Zubair Baloch. A deep semantic mobile application for thyroid cytopathology. In *Medical Imaging 2016:*

PACS and Imaging Informatics: Next Generation and Innovations, volume 9789, page 97890A. International Society for Optics and Photonics, 2016.

[KMG17a] Arkadiusz Kwasigroch, Agnieszka Mikołajczyk, and Michał Grochowski. Deep convolutional neural networks as a decision support tool in medical problems–malignant melanoma case study. In *Polish Control Conference*, pages 848–856. Springer, 2017.

[KMG17b] Arkadiusz Kwasigroch, Agnieszka Mikołajczyk, and Michał Grochowski. Deep neural networks approach to skin lesions classification—a comparative analysis. In *2017 22nd International Conference on Methods and Models in Automation and Robotics (MMAR)*, pages 1069–1074. IEEE, 2017.

[KMH+16] Hanna Källén, Jesper Molin, Anders Heyden, Claes Lundström, and Kalle Åström. Towards grading gleason score using generically trained deep convolutional neural networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1163–1167. IEEE, 2016.

[KPN+16] Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Y Ng, Pengfei Diao, Christian Igel, Celine M Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE transactions on medical imaging*, 35(5):1322–1331, 2016.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[KSQ+16] Ashnil Kumar, Pradeeba Sridar, Ann Quinton, R Krishna Kumar, Dagan Feng, Ralph Nanan, and Jinman Kim. Plane identification in fetal ultrasound images using saliency maps and convolutional neural networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 791–794. IEEE, 2016.

[KvGKdH17] Thijs Kooi, Bram van Ginneken, Nico Karssemeijer, and Ard den Heeten. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Medical physics*, 44(3):1017–1027, 2017.

[KVS+17] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017.

[LAKBM17] Avisek Lahiri, Kumar Ayush, Prabir Kumar Biswas, and Pabitra Mitra. Generative adversarial learning for reducing manual annotation in semantic segmentation on large scale miscroscopy images: Automated vessel segmentation in retinal fundus image as test case. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 42–48, 2017.

[LKB+17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[LLL+95] S-CB Lo, S-LA Lou, Jyh-Shyan Lin, Matthew T Freedman, Minze V Chien, and Seong Ki Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4):711–718, 1995.

[LLS+17] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2736–2744, 2017.

[LSY18] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[LTK16] Xinran Liu, Hamid R Tizhoosh, and Jonathan Kofman. Generating binary tags for fast medical image retrieval based on convolutional nets and radon transform. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 2872–2878. IEEE, 2016.

[LZN+18] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–34, 2018.

[MBS+14] Raphael Meier, Stefan Bauer, Johannes Slotboom, Roland Wiest, and Mauricio Reyes. Patient-specific semi-supervised learning for postoperative brain tumor segmentation. In *International Conference on Medical*

Image Computing and Computer-Assisted Intervention, pages 714–721. Springer, 2014.

[MMKSM18a] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105741M. International Society for Optics and Photonics, 2018.

[MMKSM18b] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1038–1042. IEEE, 2018.

[MPG+15] Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, Alzheimer's Disease Neuroimaging Initiative, et al. Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. *Neuroimage*, 104:398–412, 2015.

[PHS+14] Sergey M Plis, Devon R Hjelm, Ruslan Salakhutdinov, Elena A Allen, Henry J Bockholt, Jeffrey D Long, Hans J Johnson, Jane S Paulsen, Jessica A Turner, and Vince D Calhoun. Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience*, 8:229, 2014.

[PM15] Adrien Payan and Giovanni Montana. Predicting alzheimer's disease: a neuroimaging study with 3d convolutional neural networks. *arXiv preprint arXiv:1502.02506*, 2015.

[PMM+18] Sérgio Pereira, Raphael Meier, Richard McKinley, Roland Wiest, Victor Alves, Carlos A Silva, and Mauricio Reyes. Enhancing interpretability of automatically extracted machine learning features: application to a rbm-random forest system on brain lesion segmentation. *Medical image analysis*, 44:228–244, 2018.

[PPS14] Toufiq Parag, Stephen Plaza, and Louis Scheffer. Small sample learning of superpixel classifiers for em segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 389–397. Springer, 2014.

[RGS+10] J Ramírez, JM Górriz, F Segovia, R Chaves, D Salas-Gonzalez, M López, I Álvarez, and P Padilla. Computer aided diagnosis system for the alzheimer's disease based on partial least squares and random

forest spect image classification. *Neuroscience letters*, 472(2):99–103, 2010.

[RHN+17] Farhan Riaz, Ali Hassan, Rida Nisar, Mario Dinis-Ribeiro, and Miguel Tavares Coimbra. Content-adaptive region-based color texture descriptors for medical images. *IEEE journal of biomedical and health informatics*, 21(1):162–171, 2017.

[RIZ+17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[RLT+17] Alvin Rajkomar, Sneha Lingam, Andrew G Taylor, Michael Blum, and John Mongan. High-throughput classification of radiographs using deep convolutional neural networks. *Journal of digital imaging*, 30(1):95–101, 2017.

[RMC15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

[RSIK18] Alexander Rakhlin, Alexey Shvets, Vladimir Iglovikov, and Alexandr A Kalinin. Deep convolutional neural networks for breast cancer histology image analysis. In *International Conference Image Analysis and Recognition*, pages 737–744. Springer, 2018.

[SCD+17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[SCH+16] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A Helvie, Jun Wei, and Kenny Cha. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical physics*, 43(12):6654–6666, 2016.

[SCNK16] Amit Shah, Sailesh Conjeti, Nassir Navab, and Amin Katouzian. Deeply learnt hashing forests for content based image retrieval in prostate mr images. In *Medical Imaging 2016: Image Processing*, volume 9784, page 978414. International Society for Optics and Photonics, 2016.

[SCP+96] Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Datong Wei, Mark A Helvie, Dorit D Adler, and Mitchell M Goodsitt. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE transactions on Medical Imaging*, 15(5):598–610, 1996.

[SEZ+13] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[SHM+16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[SJP+11] Shantanu Singh, Firdaus Janoos, Thierry Pécot, Enrico Caserta, Gustavo Leone, Jens Rittscher, and Raghu Machiraju. Identifying nuclear phenotypes using semi-supervised metric learning. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 398–410. Springer, 2011.

[SLH+18] Thibault Sellam, Kevin Lin, Ian Yiran Huang, Michelle Yang, Carl Vondrick, and Eugene Wu. Deepbase: Deep inspection of neural networks. *arXiv preprint arXiv:1808.04486*, 2018.

[SLL07] Hyunjung Shin, Andreas Martin Lisewski, and Olivier Lichtarge. Graph sharpening plus graph integration: a synergy that improves protein functional classification. *Bioinformatics*, 23(23):3217–3224, 2007.

[SLS+14] Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *NeuroImage*, 101:569–582, 2014.

[SMB18] M Srividya, S Mohanavalli, and N Bhalaji. Behavioral modeling for mental health using machine learning algorithms. *Journal of medical systems*, 42(5):88, 2018.

[SRASC14] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[SRF16] AJ Schaumberg, MA Rubin, and TJ Fuchs. H&e-stained whole slide deep learning predicts spop mutation state in prostate cancer. biorxiv: 064279. 2016.

[SS13] Heung-Il Suk and Dinggang Shen. Deep learning-based feature representation for ad/mci classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 583–590. Springer, 2013.

[SSW+17] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

[STS+06] Hyunjung Shin, Koji Tsuda, B Schölkopf, A Zien, et al. Prediction of protein function from networks. In *Semi-supervised learning*, pages 361–376. MIT press, 2006.

[STZQ16] Wenqing Sun, Tzu-Liang Bill Tseng, Jianying Zhang, and Wei Qian. Computerized breast cancer analysis system using three stage semi-supervised learning method. *Computer methods and programs in biomedicine*, 135:77–88, 2016.

[SVD+18] Hojjat Salehinejad, Shahrokh Valaee, Tim Dowdell, Errol Colak, and Joseph Barfett. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 990–994. IEEE, 2018.

[SWC+11] William V Stoecker, Mark Wronkiewiecz, Raeed Chowdhury, R Joe Stanley, Jin Xu, Austin Bangert, Bijaya Shrestha, David A Calcara, Harold S Rabinovitz, Margaret Oliviero, et al. Detection of granularity in dermoscopy images of malignant melanoma using color and texture features. *Computerized Medical Imaging and Graphics*, 35(2):144–147, 2011.

[SYH+16] Hang Su, Zhaozheng Yin, Seungil Huh, Takeo Kanade, and Jun Zhu. Interactive cell segmentation based on active and semi-supervised learning. *IEEE transactions on medical imaging*, 35(3):762–777, 2016.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[SZL+09] Yangqiu Song, Changshui Zhang, Jianguo Lee, Fei Wang, Shiming Xiang, and Dan Zhang. Semi-supervised discriminative classification with application to tumorous tissues segmentation of mr brain images. *Pattern analysis and applications*, 12(2):99–115, 2009.

[UHE18] Hristina Uzunova, Heinz Handels, and Jan Ehrhardt. Unsupervised pathology detection in medical images using learning-based methods. In *Bildverarbeitung für die Medizin 2018*, pages 61–66. Springer, 2018.

[VGSJC15] Bram Van Ginneken, Arnaud AA Setio, Colin Jacobs, and Francesco Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International symposium on biomedical imaging (ISBI)*, pages 286–289. IEEE, 2015.

[VTBE15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[Win] NNI architecture. `https://nni.readthedocs.io/en/latest/Overview.html`. Accessed: 2019-05-17.

[WKG+16] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[WLP+14] Bo Wang, K Wei Liu, K Marcel Prastawa, Andrei Irima, Paul M Vespa, John D Van Horn, P Thomas Fletcher, and Guido Gerig. 4d active cut: An interactive tool for pathological anatomy modeling. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 529–532. IEEE, 2014.

[WPH+18] Jimmy Wu, Diondra Peck, Scott Hsieh, Vandana Dialani, Constance D Lehman, Bolei Zhou, Vasilis Syrgkanis, Lester Mackey, and Genevieve Patterson. Expert identification of visual primitives used by cnns during

mammogram classification. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105752T. International Society for Optics and Photonics, 2018.

[WPL+17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[WZP+18] Jimmy Wu, Bolei Zhou, Diondra Peck, Scott Hsieh, Vandana Dialani, Lester Mackey, and Genevieve Patterson. Deepminer: Discovering interpretable representations for mammogram classification and explanation. *arXiv preprint arXiv:1805.12323*, 2018.

[XZLL18] Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. Snas: stochastic neural architecture search. *arXiv preprint arXiv:1812.09926*, 2018.

[YLLH18] Longchuan Yan, Wantao Liu, Yin Liu, and Songlin Hu. Cgan based cloud computing server power curve generating. In *International Conference on Algorithms and Architectures for Parallel Processing*, pages 12–20. Springer, 2018.

[YWB18] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *arXiv preprint arXiv:1809.07294*, 2018.

[ZLR05] XJ Zhu, J Lafferty, and R Rosenfeld. Semi-supervised learning with graphs [ph. d. dissertation], 2005.

[ZVSL18] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

[ZXX+17] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.

[ZZM+17] Ruikai Zhang, Yali Zheng, Tony Wing Chung Mak, Ruoxi Yu, Sunny H Wong, James YW Lau, and Carmen CY Poon. Automatic detection and classification of colorectal polyps by transferring low-level cnn

features from nonmedical domain. *IEEE journal of biomedical and health informatics*, 21(1):41–47, 2017.

Appendix

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Khatia Kilanava**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Enhancing Breast Cancer Prediction Using Unlabeled Data

   supervised by Prof. Sherif Aly Ahmed Sakr and Dr. Radwa El Shawi

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Khatia Kilanava
Tartu, 21.05.2019