

WARSAW UNIVERSITY OF TECHNOLOGY

FACULTY OF MATHEMATICS AND INFORMATION SCIENCE

# Ph.D. Thesis

Mateusz Buda, M.Sc.

Computer-aided image-based diagnosis: extensions of  
computational models with domain knowledge

Supervisor

Artur Jerzy Przelaskowski, Professor

Co-supervisor

Maciej Andrzej Mazurowski, Ph.D.

WARSAW 2024



# Abstract

This dissertation presents machine learning methods to support image-based medical diagnosis. The work presents research results regarding the limitations of the effectiveness of machine learning methods in selected applications. The dissertation includes nine publications which review the concepts related to deep learning in radiology, investigate challenges in training robust machine learning models supporting image-based medical diagnosis, and describe methods to enhance their effectiveness in the context of real-world utility in important applications.

The first publication discusses general challenges and growing potential of the applications of deep learning methods and models supporting image-based diagnosis, emphasizing issues related to data availability, model validation, collaboration with clinicians, and obstacles to effectively integrate deep learning into radiology practice. The second publication investigates the class imbalance problem in training convolutional neural networks. The following publications focus on specific projects and applications that investigate these two challenges and propose solutions to overcome them.

The next two publications are devoted to applications in brain magnetic resonance imaging (MRI), where the first focuses on the segmentation of brain MRI with a highly imbalanced distribution of target classes and the second investigates the benefits of transfer learning in radiogenomics.

The following three publications are related to applications of machine learning methods in supporting thyroid diagnosis using ultrasound, where the first shows how approximate annotations can be used to extend the training data size for ultrasound images, the second describes the implementation of a system for autonomous diagnosis of thyroid nodules, and the last publication in this group describes the optimization of a guideline system for the interpretation and diagnosis of thyroid nodules in ultrasound images.

Finally, there are two publications for applications in digital breast tomosynthesis (DBT), where the first introduces a publicly released DBT dataset together with a baseline model, and the second proposes a method based on image-completion using generative adversarial networks which utilize only examples without lesions for abnormality detection.

**Keywords:** computer-aided diagnosis, machine learning, radiology, medical imaging, convolutional neural networks, deep learning, transfer learning, class imbalance, generative adversarial network, brain MRI, thyroid ultrasound, digital breast tomosynthesis, radiogenomics

## Streszczenie

Niniejsza rozprawa przedstawia metody uczenia maszynowego wspomagające obrazową diagnostykę medyczną. W rozprawie przedstawiono wyniki badań dotyczące ograniczeń skuteczności metod uczenia maszynowego w wybranych zastosowaniach, w szczególności problem ograniczonego zbioru uczącego i nierównowagi klas. Rozprawa obejmuje dziewięć publikacji, w których dokonano przeglądu koncepcji związanych z głębokim uczeniem w radiologii, zbadano wyzwania związane z uczeniem efektywnych modeli wspomagających diagnostykę obrazów medycznych oraz opisano metody pozwalające zwiększyć ich efektywność w kontekście realnej przydatności w istotnych zastosowaniach.

Pierwsza publikacja omawia ogólne wyzwania i rosnący potencjał zastosowań metod i modeli głębokiego uczenia we wspomaganiu diagnostyki obrazowej, z naciskiem na kwestie związane z dostępnością danych, walidacją modeli, współpracą z lekarzami oraz przeszkodami w efektywnym wdrażaniu uczenia głębokiego. Druga publikacja dotyczy problemu nierównowagi klas w uczeniu sieci splotowych. Kolejne publikacje koncentrują się na konkretnych projektach i zastosowaniach, które badają te dwa wyzwania i proponują rozwiązania.

Kolejne dwie publikacje poświęcone są zastosowaniom w rezonansie magnetycznym mózgu (MRI), gdzie pierwsza koncentruje się na segmentacji MRI mózgu z wysoce niezerównoważonym rozkładem klas, a druga bada korzyści płynące z transferu wiedzy w radiogenomice.

Kolejne trzy publikacje dotyczą zastosowań metod uczenia maszynowego we wspomaganiu diagnostyki tarczycy wykorzystującej ultrasonografię. Pierwsza pokazuje, w jaki sposób przybliżone anotacje można wykorzystać do rozszerzenia rozmiaru danych treningowych dla obrazów ultrasonograficznych, druga opisuje implementację systemu do autonomicznej diagnostyki guzków tarczycy, a ostatnia publikacja w tej grupie opisuje optymalizację systemu wytycznych do interpretacji i diagnostyki guzków tarczycy w obrazach ultrasonograficznych.

Dwie ostatnie publikacje dotyczą zastosowań w cyfrowej tomosyntezie piersi, z których pierwsza przedstawia publicznie udostępniony zbiór danych wraz z modelem bazowym, a druga proponuje metodę opartą na uzupełnianiu obrazu przy użyciu modelu generatywnego, który wykorzystuje tylko przykłady bez zmian chorobowych do wykrywania obszarów anormalnych.

**Słowa kluczowe:** komputerowe wspomaganie diagnozy, uczenie maszynowe, radiologia, obrazowanie medyczne, sieci splotowe, głębokie uczenie, transfer uczenia, nierównowaga klas, sieci generacyjne, rezonans mózgu, ultrasonografia tarczycy, tomosynteza piersi, radiogenomika



## Acknowledgements

First, I want to thank my supervisor, Artur Przelaskowski, for his constant help and support. His advice and encouragement have been very important to me.

I am also grateful to my co-supervisor, Maciej Mazurowski, from Duke University. He guided me in choosing the right research directions and helped me work towards impactful publications. I especially appreciate his patience with me when I was not convinced to his ideas.

My time at RAILabs was made better by all my colleagues and co-workers: Ashirbani Saha, Albert Świącicki, Nianyi Li, Yinhao Ren, Zhe Zhu, Jun Zhang, Ehab AlBadawy, Bian Harrawood, and many others. I am thankful for their teamwork and passion for research. I also want to thank all my co-authors for their contributions to our publications.

A big thank to my family and friends for their support and special thanks to my fiancé Sonia. They have always believed in me and encouraged me.

I would like to acknowledge my primary school math teacher, Iwona Guzewska, who sparked my interest in mathematics and helped start my journey.

Finally, I thank everyone who helped me in any way during my PhD. Your contributions, big or small, have been a great help in my academic and personal growth.



# Contents

<b>Chapter 1. Introduction</b>	<b>9</b>
1.1 Machine learning to support diagnostic imaging . . . . .	9
1.2 Limitations of machine learning methods supporting diagnosis in radiology . .	10
1.3 Main goals and research theses . . . . .	12
1.4 Publications comprising the thesis . . . . .	13
1.5 Thesis organization . . . . .	15
<b>Chapter 2. Challenges to applied machine learning in radiology</b>	<b>16</b>
2.1 Growing impact of artificial intelligence on radiology . . . . .	16
2.2 Class imbalance . . . . .	17
<b>Chapter 3. Applications in brain magnetic resonance imaging</b>	<b>19</b>
3.1 Brain tumor genomic subtype prediction via automated shape analysis . . . . .	19
3.2 Unbiased radiogenomic analysis with transfer learning . . . . .	20
<b>Chapter 4. Applications in thyroid ultrasound</b>	<b>23</b>
4.1 Approximate annotations for ultrasound images . . . . .	23
4.2 Incorporating auxiliary imaging feature prediction tasks for diagnosis of thyroid nodules . . . . .	24
4.3 Optimized guidelines for thyroid nodule ultrasound interpretation . . . . .	26
<b>Chapter 5. Applications in digital breast tomosynthesis</b>	<b>28</b>
5.1 Digital breast tomosynthesis screening data set for lesion detection . . . . .	28
5.2 Abnormality detection by image completion . . . . .	30
<b>Chapter 6. Summary</b>	<b>32</b>
6.1 Discussion on computer-aided diagnosis systems . . . . .	34

<b>Chapter 7. Overview of academic achievements</b>	<b>38</b>
<b>Bibliography</b>	<b>40</b>
<b>Appendix A. Publications</b>	<b>45</b>
A.1 Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI . . . . .	45
A.2 A systematic study of the class imbalance problem in convolutional neural networks . . . . .	62
A.3 Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm . . . . .	74
A.4 Deep radiogenomics of lower-grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images . . . . .	83
A.5 Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images . . . . .	91
A.6 Management of thyroid nodules seen on US images: deep learning may match performance of radiologists . . . . .	99
A.7 Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility . . . . .	107
A.8 A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images . . . . .	116
A.9 A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis .	127
<b>Appendix B. Authorship statements</b>	<b>141</b>

# Chapter 1

## Introduction

### 1.1 Machine learning to support diagnostic imaging

In the past decade, we witnessed rapid development of methods in machine learning that enabled using large amounts of labeled data. Access to computational resources helped prove the utility of ideas that appeared a long time ago but emerged as applicable only recently. A notable breakthrough happened in the image processing domain. A Deep Convolutional Neural Network (CNN), trained on the ImageNet dataset, achieved test classification error notably lower than established computer vision approaches based on human-engineered features [10]. The concept of CNN models was already introduced in Cognitron [11] and improved in the decades that followed [12–14].

In the following years, the progress in machine learning started affecting other domains, including radiology [1]. First attempts to use CNN models in medical image analysis took place in the 1990s for lung cancer detection in chest radiographs [15] as well as detection of microcalcifications [16] and masses [17] in mammography.

Computer-aided diagnosis (CAD) support systems are designed to assist clinicians in the diagnostic process. These systems can operate under different paradigms such as detection, diagnosis, staging, and treatment assessment. While CAD applications for detection aim to identify specific anomalies within imaging data, diagnosis-oriented CAD tools provide a probable cause or category for observed anomalies.

It is crucial for the intended use of a CAD system to align with the clinical environment. We can distinguish four main intended uses of a CAD system: second read, concurrent read, triage, and rule-out [18]. The second read CAD aids in decision-making by providing a secondary opin-

ion after a physician’s initial interpretation. In contrast, concurrent read CAD shows its output to the physician simultaneously with their initial reading. The triage CAD optimizes workflow by prioritizing cases, with assessments focusing on process improvements in clinical operations. Finally, the rule-out CAD implementation approach streamlines workflow by excluding normal or negative cases without clinician review. The intended use case of a CAD system affects the focus of performance evaluation, e.g., the relative importance between sensitivity and specificity.

An important element in the development and evaluation of CAD systems is the reference standard against which the system’s performance is assessed. While the most direct reference standards employ collected image data with expert annotations, these are often subjective and can vary between the experts providing them. Thus, more objective reference standards, like pathologic assessments of biopsied lesions, even though not perfect, are preferred. However, when subjective standards are unavoidable, it is best to obtain assessments from multiple experts and evaluate the variability in their conclusions.

Machine learning models presented in this thesis are applicable to computer-aided diagnosis at different stages, from basic research to validated methods, and with different intended use cases.

## **1.2 Limitations of machine learning methods supporting diagnosis in radiology**

Machine learning methods, particularly those based on deep learning, were initially developed and used for natural images. These images, whether they are everyday photographs or snapshots of the world around us, exhibit different distributions and characteristics compared to radiological images. Radiological diagnostic images, captured using various modalities like Computed Tomography (CT), Magnetic Resonance Imaging (MRI), or ultrasound (US), are significantly different in terms of their structure, texture, contrast, and the objects present in the images.

There is a growing interest in applying machine learning techniques to the field of radiology. This interest is fueled by an increasing availability of computational resources and public medical imaging datasets, which provide an opportunity to train and validate advanced machine learning models. However, without a context and deep understanding of data acquisition protocol as well as the source of annotations (reference standard), the results may have limited

applications.

Below we discuss selected limitations which are crucial for the successful application of machine learning in radiological image diagnosis.

**Data scarcity** poses a significant challenge in the application of machine learning methods for supporting diagnosis in radiology. Machine learning models, particularly based on deep learning, rely heavily on large and diverse datasets for their training. In radiology, obtaining such datasets is not straightforward. Radiological imaging studies are expensive and time-consuming, and labelling these images often requires the expertise of trained medical professionals. This problem becomes more acute when we consider specific, less common pathology where the availability of examples is naturally scarce. Data scarcity can potentially lead to overfitting, where the model performs well on the training data but fails to generalize on unseen examples.

**The class imbalance** problem is a source of another important limitation in the application of machine learning in radiology. Certain medical conditions are rare, which leads to a highly skewed distribution of classes in the training dataset. Machine learning models, when trained on such data, tend to bias their predictions towards the majority class, thereby decreasing their sensitivity to the minority class. In the context of computer-aided diagnosis, this is particularly problematic as the minority class often represents critical pathological findings. Moreover, model performance evaluation metrics need to be selected according to requirements of downstream tasks and relevant to the model intended use.

**Validation and generalization** of machine learning models supporting interpretation of diagnostic images in radiology is also challenging. These models are trained on a specific dataset, which may represent a particular patient population, and use certain imaging devices or acquisition protocols. When applied to different populations or different imaging protocols, these models may not perform as expected. This issue, known as domain shift, requires careful model validation, possibly involving multiple datasets from varied sources. In many radiological tasks and image modalities, availability of good quality and representative benchmark datasets is still an issue.

**Integration into radiology workflow** is the key to successful application of machine learning in interpretation of radiological images. It is crucial to think how computer-aided support systems can be integrated into the existing clinical workflow. This requires more than just developing machine learning models with high accuracy. We need to consider how these models will assist radiologists in their day-to-day tasks, how the model's outputs will be presented to

them, and how they will interact with the system. Will the model pre-process images, flag potential anomalies, or provide a provisional diagnosis? How will the outputs be presented to the radiologists: as a marked image, a highlighted region, or a textual report? Careful attention must also be given to downstream tasks. For instance, after the model identifies a potential issue, how will it facilitate further diagnostic procedures or treatment planning? The ultimate goal of these systems is to aid the radiologists and doctors and improve patient outcomes. Thus, their design and deployment must be carried out with a clear understanding of the radiological workflow, the needs of the radiologists, and the overall patient care process.

### 1.3 Main goals and research theses

The main objective of several years of research, the result of which is presented in this dissertation, was to improve machine learning methods to support diagnostic imaging in the context of selected, particularly relevant, challenges. The work aims to suggest solutions and refine existing approaches by incorporating specific domain knowledge in the development of a CAD system to make them work well in the context of the identified problems and considering the specific requirements of the image-based diagnosis.

The main research theses are twofold. The first one is that *it is possible to improve data-intensive machine learning methods to more effectively support image-based diagnosis* by taking into account domain-specific considerations and requirements (knowledge models, assessment procedures, forms of assessment) and effectively addressing the problems of insufficient training data and class imbalance. The preferred method was to explore the application of machine learning methods to computer-aided diagnosis, where training data is often limited compared to the complexity of the problem being solved and the distribution of target classes is imbalanced. The results suggest that by addressing these challenges, machine learning methods can be effectively applied to specific diagnostic problems, resulting in improved accuracy and efficiency of the diagnostic process.

The second research thesis is that *it is possible to effectively incorporate machine learning methods into the assessment workflow to facilitate diagnosis*. This concept aims to investigate the feasibility and effectiveness of integrating machine learning methods into the diagnostic process, taking into account the specific requirements and constraints of the radiology workflow. Thus, by developing solutions tailored to the specific domain, they can be integrated into radi-



ologists' workflows and the accuracy and efficiency of radiological diagnoses can be improved. The aim was to develop and validate machine learning methods that can be integrated into the radiology workflow and can be used by radiologists to support their diagnostic decisions.

A complementary list of more specific research hypotheses is as follows:

- Class imbalance, ubiquitous in medical imaging, negatively affects the performance of deep convolutional neural networks.
- Oversampling is an effective method to train a brain MRI segmentation U-net model for estimating tumor shape features.
- Transfer learning from a similar domain improves the results of convolutional neural networks.
- Approximate annotations for ultrasound images can help extend the training dataset size and improve the performance of a deep learning-based segmentation model.
- Auxiliary annotations relevant for diagnosis can be utilized for multi-task learning of a model for interpretation of thyroid nodule ultrasound images.
- A data-driven approach can help optimize and simplify a guideline system for the management of thyroid nodules.
- A public benchmark dataset and baseline model are important for validating and comparing the performance of machine learning methods.
- A training dataset containing only negative examples (not containing lesions) can be utilized for the development of a method for abnormality detection based on generative adversarial networks.

## 1.4 Publications comprising the thesis

The thesis is comprised of the nine scientific articles listed in this section, all published in peer-reviewed international journals. Full text of each article is available in the Appendix [A](#).

[[A.1](#)] Maciej A Mazurowski, **Mateusz Buda**, Ashirbani Saha, and Mustafa R Bashir. Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. *Journal of Magnetic Resonance Imaging*, 49(4):939–954, 2019

- [A.2] **Mateusz Buda**, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018
- [A.3] **Mateusz Buda**, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology & Medicine*, 109:218–225, 2019
- [A.4] **Mateusz Buda**, Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep radiogenomics of lower-grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images. *Radiology: Artificial Intelligence*, 2(1), 2020
- [A.5] **Mateusz Buda**, Benjamin Wildman-Tobriner, Kerry Castor, Jenny K Hoang, and Maciej A Mazurowski. Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images. *Ultrasound in Medicine & Biology*, 46(2):415–421, 2020
- [A.6] **Mateusz Buda**, Benjamin Wildman-Tobriner, Jenny K Hoang, David Thayer, Franklin N Tessler, William D Middleton, and Maciej A Mazurowski. Management of thyroid nodules seen on US images: deep learning may match performance of radiologists. *Radiology*, 292(3):695–701, 2019
- [A.7] Benjamin Wildman-Tobriner, **Mateusz Buda**, Jenny K Hoang, William D Middleton, David Thayer, Ryan G Short, Franklin N Tessler, and Maciej A Mazurowski. Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility. *Radiology*, 292(1):112–119, 2019
- [A.8] **Mateusz Buda**, Ashirbani Saha, Ruth Walsh, Sujata Ghate, Nianyi Li, Albert Świącicki, Joseph Y Lo, and Maciej A Mazurowski. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. *JAMA network open*, 4(8), 2021
- [A.9] Albert Swiecicki, Nicholas Konz, **Mateusz Buda**, and Maciej A Mazurowski. A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis. *Scientific reports*, 11(1):1–13, 2021

## 1.5 Thesis organization

The reminder of the thesis is organized as follows. Chapter 2 examines the challenges related to the development of machine learning methods for medical imaging in general, and for radiology in particular. It describes the challenges related to the limited and imbalanced data, as well as the specific requirements and constraints in the context of the radiology domain. The following three chapters present projects that address the challenges described in Chapter 2 in selected applications. Chapter 3 is focused on applications in brain magnetic resonance imaging, while Chapter 4 describes applications in thyroid ultrasound. Chapter 5 examines applications in digital breast tomosynthesis.

Chapter 6 provides a summary of the thesis followed by a discussion on computer-aided diagnosis systems whereas Chapter 7 gives an overview of the academic achievements of the thesis author. Finally, two appendices are included at the end of the thesis. Appendix A contains copies of the full text of publications comprising the thesis, while Appendix B provides authorship statements from co-authors.

## Chapter 2

# Challenges to applied machine learning in radiology

### 2.1 Growing impact of artificial intelligence on radiology

The impact of deep learning on radiology is growing, as it demonstrates exceptional performance in image analysis tasks. This technology offers potential improvements in various aspects of radiology, including disease detection, diagnosis, characterization, and workflow efficiency. Despite these promising advancements, several challenges remain, such as data availability, model overfitting, and proper validation of algorithms for clinical use relevant to the CAD intended use case [1].

Recent success in deep learning can be attributed to the availability of large datasets, increased processing power, and rapid algorithmic development. These factors have contributed to the emergence of deep learning applications in radiology, overcoming initial inertia due to the need for medical imaging expertise and limited availability of large medical imaging datasets [8].

Several challenges and pitfalls must be addressed to effectively integrate deep learning into radiology practice. The availability of data is a significant challenge, as medical imaging datasets are often smaller than those used for natural images. This limitation, combined with the large number of parameters in deep neural networks, increases the risk of overfitting and potentially reduces model performance on new data. To mitigate this issue, researchers can pre-train models with other datasets, use smaller models, or augment data with slight alterations of original images.

Another challenge is the proper validation of developed deep learning models in the context

of the CAD intended clinical use case. This requires posing clinically significant questions, careful curation of datasets, and precise definitions of non-imaging variables such as pathology, genomic markers, and patient outcomes. Close collaboration with clinicians and other experts is crucial at various stages of development.

The future of deep learning in radiology presents multiple challenges. Technological hurdles need to be overcome to demonstrate that deep learning algorithms can replace or augment radiologists' work. Legal and ethical challenges, such as determining responsibility for algorithmic errors, need to be addressed. Patient acceptance of human-free image interpretation and regulatory issues will also play crucial roles. Finally, practical concerns about incorporating deep learning algorithms into radiology workflow without disrupting the practice must be resolved.

In the paper [A.1] *Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI*, the author made notable contributions by drafting and reviewing the manuscript, providing a comprehensive introduction to deep learning methods, and conducting a thorough review of the state of the art in classification and segmentation in radiology. Additionally, the author prepared visual materials to enhance the clarity and impact of the publication, making it more accessible and engaging for readers.

## 2.2 Class imbalance

Class imbalance is a problem connected to insufficient training data and refers to a skewed distribution of classes in a dataset [2]. It is especially pronounced in medical datasets, since class distribution reflects the natural prevalence of diseases in screening populations [1]. Working with imbalanced datasets requires carrying out a model evaluation with performance metrics robust to uneven class distribution.

In a comprehensive experimental study of the class imbalance problem in training CNN models, we used benchmark datasets and established network architectures to compare some commonly used methods, i.e., oversampling, undersampling, two-phase training, and thresholding that compensates for prior class probabilities. Based on our experiments, we concluded that (i) class imbalance has a detrimental effect; (ii) oversampling emerged as the dominant method; (iii) oversampling should be applied to the extent that eliminates the imbalance, whereas the optimal undersampling ratio depends on the extent of imbalance; (iv) oversampling does not lead to overfitting of CNNs; (v) thresholding should be applied when accuracy is of interest.

Other studies explored adaptive weighting of examples by design of a loss function. Focal loss is a variant of cross-entropy that diminishes the contribution of almost correctly classified cases [19]. This way, easy to classify inputs from majority classes do not overwhelm model weights' updates and give a chance for improvement to examples from minority classes that produce highly incorrect predictions.

For the publication [A.2] *A systematic study of the class imbalance problem in convolutional neural networks*, the author played a significant role in all stages of this project, including formulating research goals, conducting a literature review, and developing the methodology and experimental design. Additionally, he developed the software to run experiments, collected evidence, and analyzed the results. The source code for All-CNN architecture, developed by the author, used in the experiments is shared in a public repository <sup>1</sup>. The author of this thesis wrote the manuscript, formulated conclusions, and prepared visualizations and plots to support the findings.

---

<sup>1</sup><https://github.com/mateuszbuda/ALL-CNN>

## Chapter 3

# Applications in brain magnetic resonance imaging

### 3.1 Brain tumor genomic subtype prediction via automated shape analysis

Machine learning can assist radiologists in labour-intensive tasks. In many cases, promising results for the prognostic value of various quantitative imaging features are challenging to apply in practice, since radiologists would be required to segment abnormal areas on multiple images manually [20]. Moreover, this method suffers from both inter-reader and intra-reader variability that may significantly affect predicted outcome [21]. Outsourcing delineation of lesions to an automated system offers consistency and saves time. However, an algorithm performing this task is expected to deliver human-level quality of predicted segmentations.

Using a public dataset of magnetic resonance images (MRI) for 110 patients from 5 institutions, we trained a CNN with U-Net architecture to perform segmentation of brain tumors [3]. The tumors in brain MRI are considerably smaller compared to the background (air) and normal brain tissue. Based on insights from a study on the class imbalance in training CNNs, we performed undersampling by discarding images that did not contain any brain tissue, oversampled images of tumors, and applied data augmentation.

The model achieved 82% mean Dice coefficient, which is comparable to human performance. Next, predicted segmentation volumes were validated in genomic analysis of tumor shape features previously discovered to be prognostic of patient outcomes. Strong associations

were found between clinically relevant genomic subtypes and tumor shape features that were extracted in a fully automatic way.

This method allows for obtaining estimated characteristics of tumor genomics in a non-invasive way that does not require extra effort from radiologists to annotate MRI sequences manually. Even imprecise information of tumor genomics from MRI, accessible in the early stages of treatment, could be used to guide therapy until more accurate genomic test results are available.

In the paper [A.3] *Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm*, the author made significant contributions to various aspects of the research. He was involved in conceptualizing the research goals and designing the methodology. He developed the software and implemented the U-Net architecture for the segmentation of brain tumors in MRIs which was shared in a public repository<sup>1</sup> together with the pre-processed dataset used in the project to ensure reproducibility<sup>2</sup>. The author also played a crucial role in performing formal analysis and investigation of the results. His contributions also included data curation. The author of this thesis participated in writing the original draft, reviewing, and editing the manuscript.

## 3.2 Unbiased radiogenomic analysis with transfer learning

Radiogenomics is an area of cancer research that explores the relationship between imaging characteristics of a lesion and its gene expression patterns or mutations [22]. CNN models excel in learning hierarchical feature extractors that reflect statistics of patterns present in the training data. This property can be utilized in radiogenomic analysis to reduce human bias related to identifying imaging-based features predictive of tumor genomics.

In a recent study [4], we investigated the impact of a domain gap between datasets used for pre-training and fine-tuning of a CNN. The target task used for evaluation was classifying MR images of gliomas (brain tumors) to genomic subtypes. The study involved open access imaging and genomic data. We tested the effect of pre-training a CNN on natural images (ImageNet dataset [23]) and brain MRI containing a similar type of tumor. As a baseline, we trained a CNN from random initialization of weights (without pre-training on external data). Following the findings from previous studies [20], input images contained extracted tumors, as shown in

---

<sup>1</sup><https://github.com/mateuszbuda/brain-segmentation-pytorch>

<sup>2</sup><https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation>



Figure 3.1, to provide shape information predictive of genomic subtypes to the models.

The best performing method for the task of discriminating between tumor genomic subtypes that show significantly different survival times was transfer learning utilizing an MRI dataset for pre-training, with the area under the receiver operating characteristic curve (AUC) of 0.73. In comparison, for the networks trained from scratch and pre-trained on natural images, the AUC was at 0.68 and 0.64, respectively.

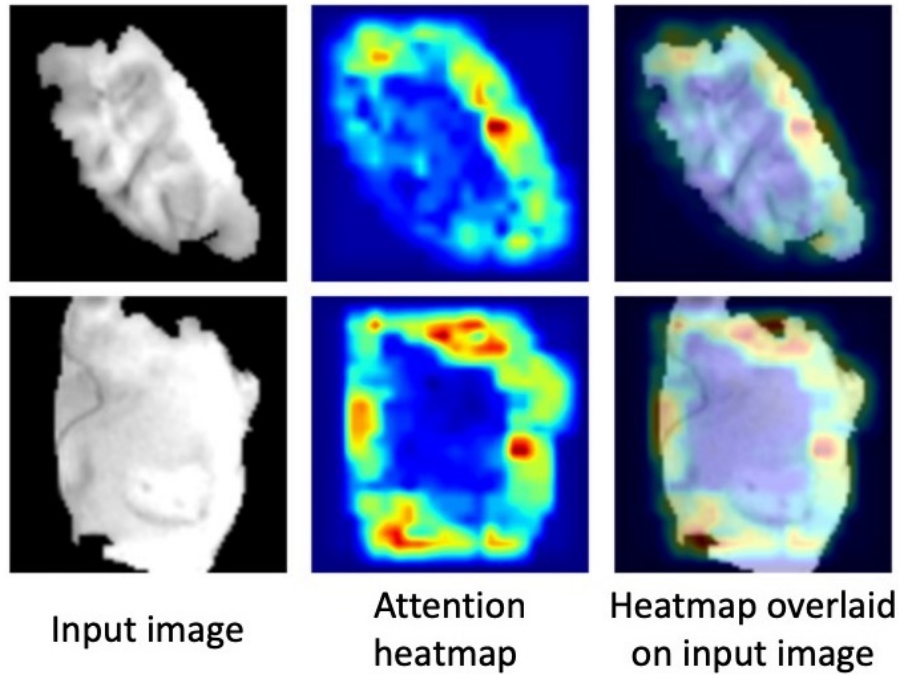


Figure 3.1: CNN attention heatmaps indicate areas contributing to prediction [4].

Although the imaging-based methods are not yet ready to replace genomic testing, they provide valuable information that can facilitate patient treatment in various ways. Furthermore, results from genomic tests are, to some extent, sensitive to tissue sampling. Areas of the tumor that were identified to contribute to the model prediction could be used to guide biopsies for more accurate and reliable outcomes (Figure 3.1).

In the publication [A.4] *Deep radiogenomics of lower-grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images*, the author made substantial contributions to various aspects of the research. He conceptualized the research goals and developed the methodology for exploring transfer learning in radiogenomics. The author was responsible for developing the software and implementing the CNN models for classifying gliomas based on genomic subtypes. He also performed validation, formal analysis, and investigation of the results, ensuring the accuracy and reliability of the findings. Python code used for statistical

comparison of tested methods using bootstrapping was shared in a public repository <sup>3</sup>. The author actively managed data curation and took part in writing the original draft, as well as reviewing and editing the manuscript.

---

<sup>3</sup><https://github.com/mateuszbuda/ml-stat-util>

# Chapter 4

## Applications in thyroid ultrasound

### 4.1 Approximate annotations for ultrasound images

Deep neural networks are robust to noise in training data. In the context of medical imaging, it is often the case that approximate or imprecise labels are easily available. These approximate labels can be obtained by automatic inference of labels from available radiological reports using natural language processing tools or retrieved with image processing algorithms from burned-in data present in the images [5]. Incorporating approximate annotations can notably expand training data size for supervised learning methods.

Ultrasound (US) images contain measurement markers enclosing regions of interest placed on images by technicians (Figure 4.1). The markers can be detected and used to automatically generate approximate segmentation masks for a training set with many examples without additional input from human annotators. In our research, we applied this approach to the segmentation of thyroid nodules in ultrasound images and compared it to training with precise outlines provided by an expert radiologist [5].

The segmentation model trained on 2156 US images with automatically generated annotations achieved 85% test Dice similarity coefficient (DSC) as compared to 90% DSC for the model trained on the same cases but based on outlines provided by experts. However, when 20% of images with precise segmentation masks were used for training, simulating limited resources for data collection, the performance degraded to 85%, matching that of the model trained on automatically generated annotations.

Object semantic segmentation in US images has many potential applications for downstream tasks discussed in the following sections. It can be used, for example, in object shape analysis.

However, the primary application is identification of the region of interest for further processing like diagnostic evaluation or extraction of other information relevant to diagnosis or treatment.

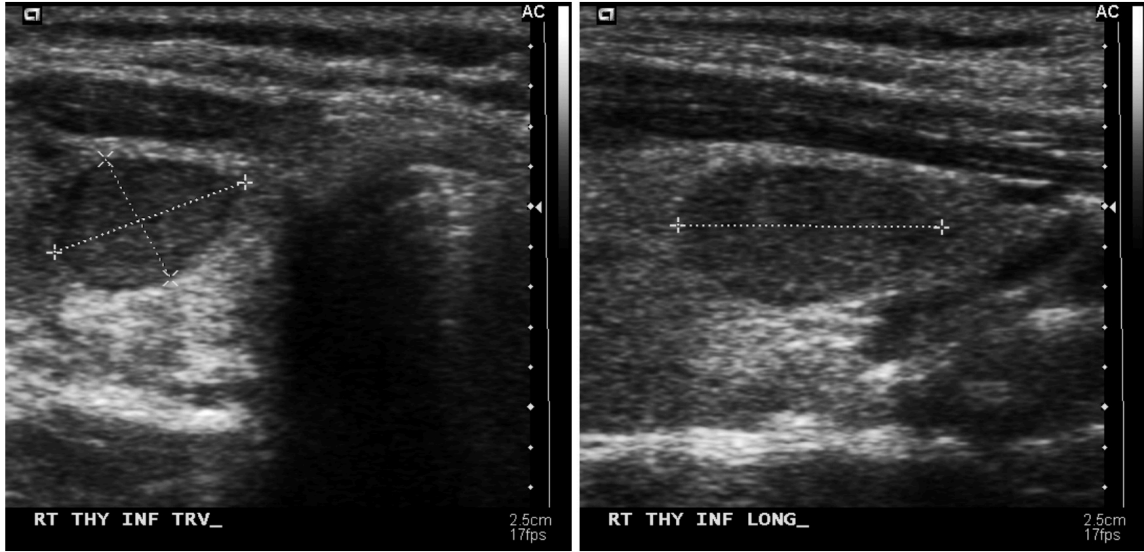


Figure 4.1: Example ultrasound images of a thyroid nodule with caliper marks. [5]

In the publication [A.5] *Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images*, the thesis author was involved in the idea development and the design of the research methodology for utilizing approximate annotations in the segmentation of thyroid nodules in ultrasound images. He was also the main contributor to the creation and implementation of the software used to produce and analyse the results. The code for the detection model was shared in public repository<sup>1</sup>. The author ensured the validity of the results by conducting thorough validation and formal analysis, while also taking part in the investigation process. Furthermore, he was responsible for managing data curation and actively participated in the writing of the original draft and contributing to its review and editing.

## 4.2 Incorporating auxiliary imaging feature prediction tasks for diagnosis of thyroid nodules

This section presents a computer-aided diagnosis system developed for a rule-out use case. We developed and evaluated a computer-aided diagnosis system for managing thyroid nodules that matches the performance of expert radiologists [6]. We trained a CNN model to distinguish

<sup>1</sup><https://github.com/mateuszbuda/deep-thyroid-nodules>

between benign and malignant lesions detected in ultrasound images. We based our system development and evaluation on a dataset of 1230 patients. The final diagnosis was obtained from biopsy results and could not be obtained merely from radiological images. Given the fact that we worked on a small dataset with a high class imbalance, we applied data augmentation with oversampling to obtain a uniform target class distribution and optimized the model using the focal loss function.

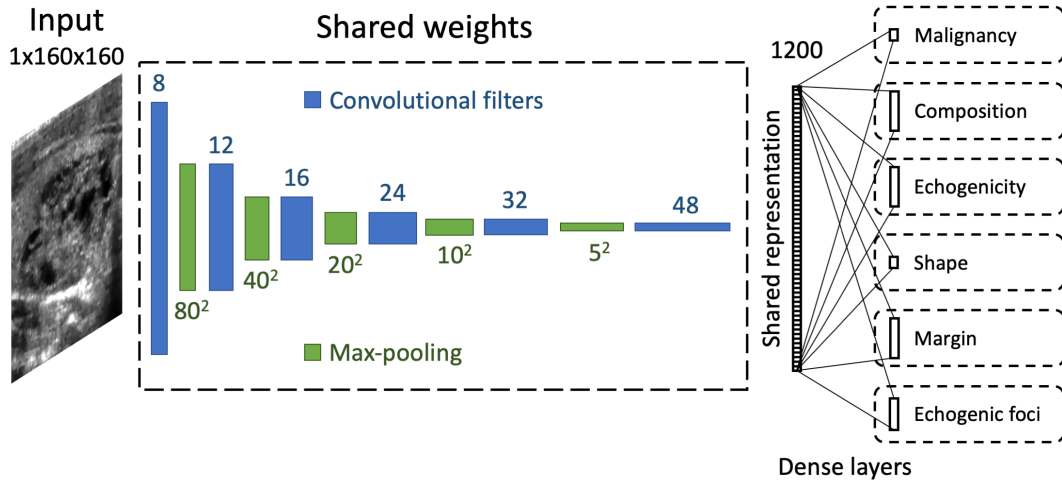


Figure 4.2: Architecture of a multi-task deep convolutional neural network trained for classification of thyroid nodules based on ultrasound images [6].

When interpreting thyroid nodules on ultrasound images, radiologists report visual features across five categories that are predictive of malignancy. They include nodule composition, echogenicity, shape, margin, and calcifications. We utilized them as additional labels for auxiliary tasks in a multi-task learning setting (Figure 4.2). By sharing network weights responsible for feature extraction, we encouraged the model to produce a shared representation of the input image relevant to all tasks, preventing overfitting and improving generalization. On the test set of 99 cases, we achieved an AUC of 0.87 compared with 0.91 obtained by consensus of three expert radiologists ( $p=0.41$ ) and 0.82 mean AUC of nine radiologists ( $p=0.38$ ).

The model’s performance was subsequently validated on a novel dataset comprised of 378 thyroid nodules from 320 patients. This data was collected from an institution different from the one that provided the original training and test images, and it also included ultrasound (US) images acquired from new device types [24]. A reader study involving four radiologists confirmed that the model’s performance with AUC of 0.69 was again comparable to that of the radiologists, who had AUC values of 0.63, 0.66, 0.65, and 0.63, respectively.

The thesis author actively contributed to the research presented in publication [A.6] *Management of thyroid nodules seen on US images: deep learning may match performance of radiologists*. He played a key role in conceptualizing the idea of an automated diagnosis system for thyroid nodules and designing the methodology to address the inherent challenges. The author also developed the necessary software, ensuring its functionality and effectiveness. The code for the developed multi-task neural network classification model was shared on a public repository <sup>2</sup>. In addition, the author carried out the essential validation and formal analysis of the results, while participating in the research investigation and handling data curation. He was involved in the writing process of the original draft and participating in its review and editing.

### 4.3 Optimized guidelines for thyroid nodule ultrasound interpretation

Computer-aided diagnosis systems introduce a significant change to the clinical workflow. They are often evaluated and compared to human readers in their performance of a specific task, e.g. cancer detection based on one image modality. In their daily routine, clinicians take much more complex decisions after reviewing a patient’s history. For example, clinicians may recommend a follow-up after some time, order additional imaging, or perform a biopsy. Incorporating a (possibly) very accurate system that provides extra input in one step on the patient’s journey is not a trivial task.

By using genetic algorithms, we helped radiologists optimize guidelines for interpretation of thyroid nodules in ultrasound images [7]. Most importantly, this did not require any change in their workflow as they were already using a similar interpretation guideline system.

In ACR TI-RADS, radiologists assign points to a nodule based on the presence of imaging features and then use a rule-based system to infer a recommended decision. The purpose is to reduce variability among various readers as evaluation of visual features on images is less subjective than biopsy recommendation.

Using a dataset of 1425 nodules with biopsy-proven diagnoses, we optimized the assignment of points for different imaging feature categories to create the AI TI-RADS system that improved specificity in recommending biopsy. In addition, the assignment of points was simplified, which might contribute to broader adoption and, as a result, help address the problem of over-diagnosis

---

<sup>2</sup><https://github.com/MaciejMazurowski/thyroid-us>

of thyroid nodules. Our findings were validated and confirmed by independent studies that applied AI TI-RADS to cases collected from various institutions and countries [25–27].

The author was actively involved in the research for the publication [A.7] *Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility*. His contributions spanned multiple aspects of the project, including the conceptualization and methodology development for optimizing guidelines for thyroid nodule interpretation using genetic algorithms. He was responsible for software development, validation, and formal analysis of the results, in addition to investigation and data curation. The author played an important role in drafting and revising the manuscript and creating visualizations to effectively convey the research findings. Showcasing a commitment to open research, the author shared the code for the developed model on a public repository <sup>3</sup> and provided access to a web application for the optimized AI-TRIADS system <sup>4</sup>.

---

<sup>3</sup><https://github.com/mateuszbuda/AI-TI-RADS>

<sup>4</sup><https://deckard.duhs.duke.edu/ai-ti-rads>

# Chapter 5

## Applications in digital breast tomosynthesis

### 5.1 Digital breast tomosynthesis screening data set for lesion detection

The scarcity of publicly available data for training and evaluating machine learning models in radiology, and digital breast tomosynthesis (DBT) in particular, presents significant challenges. Access to well-curated and annotated medical data is limited due to the lower availability of medical images compared to natural images, strict policies governing data sharing, and the time-consuming process of deidentification and compliance. Furthermore, the annotation of medical imaging data typically requires the expertise of radiologists who are already in high demand.

To address these challenges, we curated and annotated a data set consisting of over 22,000 three-dimensional (3D) DBT volumes from 5,060 patients [8]. DBT is a relatively new modality for breast cancer screening that uses multiple cross-sectional slices for each breast, providing better performance than traditional projection images. We have made the data set publicly available at the Cancer Imaging Archive, allowing researchers to improve their algorithms and test them on the same data set, leading to higher-quality models and better comparisons between algorithms.

In addition to sharing the data set, we developed a single-phase deep learning model for detecting abnormalities in DBT and made it publicly available. This baseline model can be used for fine-tuning or solving other medical imaging tasks. The limited number of positive



locations presented a significant challenge in developing the model, but various loss functions were evaluated and compared to address this issue.

The curated DBT data set and baseline model will enable researchers to develop and validate artificial intelligence (AI) tools more effectively, potentially leading to significant advances in the field of radiology. By making this data set publicly available, the study promotes transparency, reproducibility, and collaboration in the development of AI algorithms for radiology applications.

An important factor contributing to the usefulness and clinical applicability of this curated data set is its representative prevalence of different study groups comprising (i) cancerous and (ii) benign lesions, (iii) actionable studies which were suspicious enough that resulted in additional imaging, and (iv) the largest group of normal cases. In contrast to other breast cancer dataset which focus only on examples with lesions, in our case, the prevalence of different groups of patients is similar to real world screening population. This allows for more realistic evaluation of machine learning models which is meant to assess the ability of finding a relatively small number of cancerous lesions among much larger group of normal cases.

In the publication [A.8] *A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images*, the author played a vital role in numerous aspects of the research. He made significant contributions to the conceptualization and methodology. One of his key contributions was curating and annotating the shared DBT dataset consisting of over 22,000 3D volumes from 5,060 patients, making it publicly available and easy to utilize for fellow researchers. The author also took part in software development, validation, and formal analysis, as well as the investigation process. He was responsible for writing the original draft, participating in the review and editing process. The author developed the software for working with the data, shared in a public code repository <sup>1</sup>, together with the baseline detection model <sup>2</sup>. His efforts in data curation and sharing have paved the way for advancements in the field of radiology, enabling researchers to build upon this work more effectively [28].

---

<sup>1</sup><https://github.com/mazurowski-lab/duke-dbt-data>

<sup>2</sup><https://github.com/mateuszbuda/duke-dbt-detection>

## 5.2 Abnormality detection by image completion

For some diseases, e.g. breast cancer, the prevalence of positive cases in a screening population is lower than 1%. Therefore, training a supervised lesion detection model for tasks like these may be challenging. Despite higher availability of images without abnormalities, current state-of-the-art methods rely mainly on examples containing lesions. An additional shortcoming of such a supervised approach is that it only works on types of abnormalities available in the training data and might be unreliable when tested on new rare cases.

We explored a fundamentally different method for abnormality detection in radiology that utilized only normal images for training [9]. Radiologists learn to recognize the visual structure of normal tissue and can identify abnormalities in places where the tissue is different from the expected norm. We developed an algorithm that mimics this approach and tested it on a publicly available dataset of digital breast tomosynthesis images [8].

The main component of our algorithm was a generative adversarial network (GAN) trained to perform image completion. We repeatedly removed parts of the image by applying a sliding window and inpainted them using GAN trained on normal tissues (Figure 5.1). Then, we computed reconstruction errors for all patches and combined them into image abnormality heatmap as shown in Figure 5.2. Our hypothesis was that for parts of the image presenting abnormalities, the completion network will inpaint a normal-looking tissue structure that is very different from the abnormality and, therefore, yield a high error value on the abnormality heatmap.

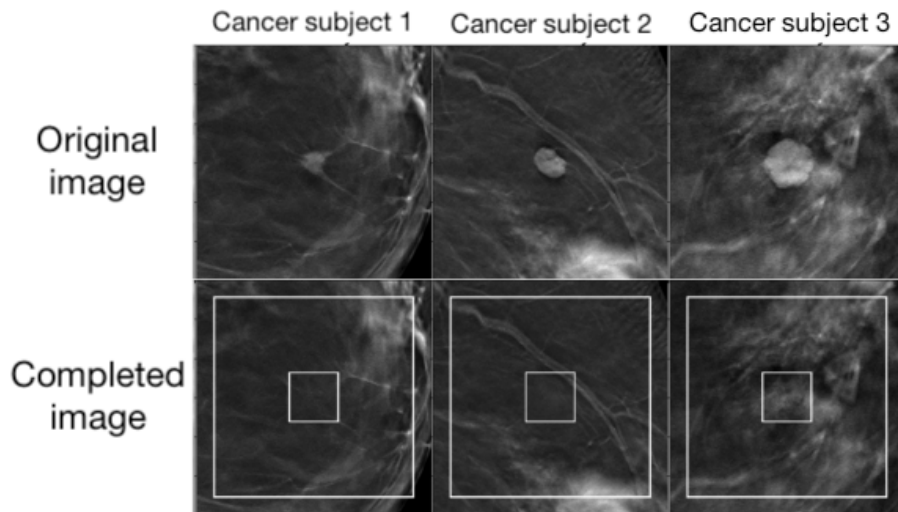


Figure 5.1: Image completion results for patches containing cancerous masses from three representative subjects. [9]

The algorithm was tested on 70 images containing bounding boxes for cancerous lesions placed by radiologists. The ability to detect abnormal tissue was validated by comparing reconstruction errors inside and outside of the bounding boxes. The mean ratio between heatmap values inside and outside of the ground truth bounding boxes was 2.77, with standard deviation of 1.79 across the test cases. This clearly indicates that the generated heatmaps indicated abnormal tissue in the images.

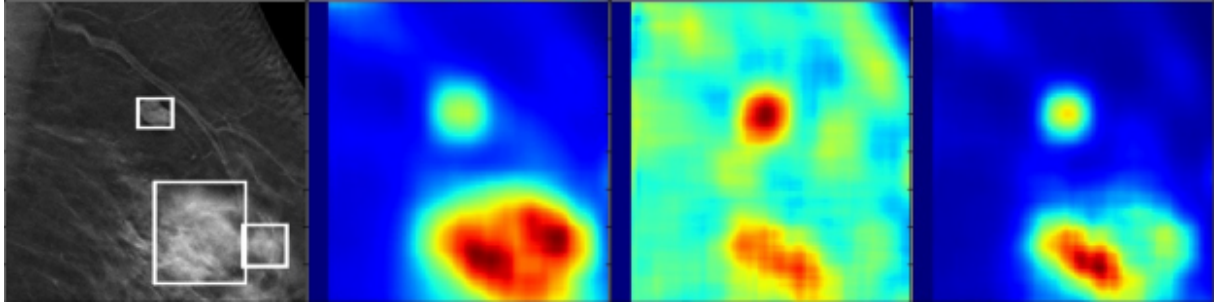


Figure 5.2: Heatmaps for a patient with cancerous masses. [9]

Our method achieved promising results, however, it was tested on a limited number of cases. Also, it has a significant computational footprint due to thousands of GAN model inferences required per image. Moreover, reconstruction error heatmaps require additional post-processing and false positive reduction for comparison with detection systems using evaluation metrics like sensitivity or free-response receiver operating characteristic curve.

In the publication [A.9] *A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis*, the author made substantial contributions to various aspects of the research. He was involved in conceptualizing the idea of using a fundamentally different method for abnormality detection in DBT which utilizes only normal images for training. He also contributed to the design of the methodology and development of the algorithm. The author participated in the software development, validation, formal analysis, and investigation of the generative adversarial network (GAN) used in the study and curated the dataset of digital breast tomosynthesis images used for training and testing. The author also contributed to writing the original draft and participated in the review and editing process of the published manuscript.

# Chapter 6

## Summary

In summary, this thesis presents significant contributions to the research in machine learning applied to radiology, with a focus on the refinement and evaluation of CAD algorithms and deep learning models, as well as the creation and sharing of publicly available datasets and code to promote collaboration and advancement in the field.

The first publication, "Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI", discusses the growing impact of artificial intelligence on radiology. Deep learning demonstrates exceptional performance in image analysis tasks, offering potential improvements in disease detection, diagnosis, characterization, and workflow efficiency. However, the author highlights several challenges, such as data availability, model overtraining, and proper validation of algorithms for clinical use. The publication also addresses the future of deep learning in radiology, emphasizing the need to overcome technological, legal, ethical, and practical challenges in order to effectively integrate deep learning into radiology practice.

The second publication, "A systematic study of the class imbalance problem in convolutional neural networks", addresses the issue of class imbalance, which is particularly pronounced in medical datasets. This problem arises from the skewed distribution of classes in a dataset, reflecting the natural prevalence of diseases in screening populations. The author conducted a comprehensive experimental study to compare commonly used methods for dealing with class imbalance, such as oversampling, undersampling, two-phase training, and thresholding. The study's conclusions offer valuable insights into the effects of class imbalance on CNN models and suggest that oversampling is the dominant method, among tested approaches, for mitigating these effects.

The third publication, "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm", investigates the potential of machine learning to assist radiologists with labor-intensive task of manual segmentation of abnormal areas. The author utilized a public dataset of magnetic resonance images (MRI) to train a U-Net model to perform segmentation of brain tumors. By addressing the class imbalance issue, the model achieved a Dice coefficient comparable to human performance. The authors then validated the predicted segmentation volumes in genomic analysis, discovering strong associations between clinically relevant genomic subtypes and tumor shape features extracted automatically. This non-invasive method offers valuable insights into tumor genomics in the early stages of treatment, potentially guiding therapy until more accurate genomic test results become available, without requiring extra manual effort from radiologists.

The fourth publication, "Deep radiogenomics of lower-grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images", delves more into the field of radiogenomics, which examines the relationship between imaging characteristics of a lesion and its gene expression patterns or mutations. The authors demonstrated the effectiveness of CNN models in learning hierarchical feature extractors that reflect patterns present in training data, thus reducing human bias in identifying imaging-based features predictive of tumor genomics. In the study, the authors investigated the impact of a domain gap between datasets used for pre-training and fine-tuning a CNN, focusing on classifying MR images of gliomas (brain tumors) to genomic subtypes.

The publication "Management of thyroid nodules seen on US images: deep learning may match performance of radiologists" presents the development and evaluation of a computer-aided diagnosis system for managing thyroid nodules. The authors addressed challenges associated with small and imbalanced datasets by applying data augmentation with oversampling and optimizing the model with focal loss. Additionally, lesions' visual features predictive of malignancy were incorporated as additional labels in a multi-task learning setting, which improved generalization and prevented overfitting. The developed system closely matched the performance of expert radiologists, demonstrating the potential of deep learning in aiding the management of thyroid nodules based on ultrasound images.

The publication "Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility" demonstrates how genetic algorithms can optimize guidelines for interpreting thyroid nodules in ultrasound images without altering radiolo-

gists' workflow. The researchers improved the ACR TI-RADS system by using AI methods to optimize the assignment of points for different imaging feature categories, leading to the creation of the AI TI-RADS system. This improved specificity in recommending biopsies and simplified the point assignment process, potentially contributing to broader adoption and addressing the issue of thyroid nodule over-diagnosis. The findings were validated by independent studies across various institutions and countries, showcasing the potential of AI in enhancing radiologists' decision-making processes.

The publication "A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images" deals with challenges in detection of lesions in digital breast tomosynthesis (DBT). The authors curated and annotated a dataset of over 22,000 3D DBT volumes from 5,060 patients, made it publicly available, and developed a baseline single-phase deep learning model for detecting abnormalities. This dataset and model will help researchers in radiology and medical imaging to develop and validate AI tools more effectively.

Finally, the last publication "A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis" proposes a fundamentally different method for abnormality detection in DBT using only normal images for training. The authors developed an algorithm based on generative adversarial network trained to perform image completion, which generated abnormality heatmaps. Despite promising results, the method was tested on a limited number of cases and had a significant computational footprint. The reconstruction error heatmaps also require further post-processing and false positive reduction for comparison with other detection systems.

## **6.1 Discussion on computer-aided diagnosis systems**

Developing a computer-aided diagnosis (CAD) support system involves several key steps to ensure its effectiveness in real-world clinical settings. First, it is crucial to decide what the CAD system will be used for by understanding the specific clinical needs. Next, the necessary data is gathered together with the reference standard relevant to the intended use case. After that, the actual algorithm or machine learning model is developed and tested. The final and most important step is actual assessment of the effects of the CAD system on patients.

The intended use of a CAD system refers to the specific clinical need it aims to address. The

system must fit into the clinical environment, characterized by factors like the patient population, the type of imaging devices used, the specific diagnostic task, and the stage in the patient care journey, e.g., screening, detection, staging, treatment assessment, or follow-up.

Additionally, there are distinct paradigms for CAD system use cases, such as second read, concurrent read, triage, and rule-out [18]. The second read CAD gives a secondary review after the initial interpretation by a physician, who does not directly interact with the CAD and this paradigm least affects the clinical workflow. On the other hand, the concurrent read offers real-time assistance, presenting the CAD output to the physician at the same time as the initial interpretation. This poses a risk of physician either over-relying or ignoring the CAD output. The triage CAD focuses on workflow efficiency, where all cases are interpreted, but the order in which they are assessed can be prioritized by the CAD system. This paradigm can also be combined with concurrent read to provide explanation on case priority. Finally, CAD systems with performance close to human experts may independently handle certain tasks or cases and be used in rule-out paradigm to identify cases with no issues, without the need for human review.

The data used for training, validation, and testing of a CAD system should mirror the desired use case and population demographics, ensuring results are replicable in genuine clinical scenarios. Flawed data collection can introduce biases and potentially lead to misinterpretation of model performance. Therefore, detailed documentation of the data collection, including selection criteria and patient demographics, is essential.

Best practices suggest consecutively sampled cases across multiple sites within defined time-frame [29]. Stratified sampling can be preferred when dealing with extremely low prevalence of a disease to ensure decent representation needed for model training. Since convenience samples are prevalent in initial studies, any conclusions drawn from them require cautious interpretation due to potential limited generalizability. A significant challenge is that many machine learning models, when trained on data from a single site, fail to generalize when applied elsewhere.

Many research groups independently gather data, leading to datasets that might lack comprehensive diversity. Public image datasets address this by offering universally accessible data repositories. However, creating such datasets involves rigorous quality checks and data de-identification. These public resources are invaluable for the progress of machine learning in medical imaging, though users must recognize and work within their inherent limitations.

The reference standard, annotations in medical images or assigned labels, is fundamentally different compared to natural images or other machine learning application domains and need

to be taken into account in the development of a CAD system.

First key consideration is related to subjective and objective annotations. Subjective annotations are primarily based on expert opinions. They are susceptible to variability, especially when based on individual interpretations. Although they gain in reliability when consensus among multiple experts is used, inherent variability remains a challenge. Subjective reference standards, especially those from multiple domain experts, are undoubtedly valuable but can be costly to obtain for large datasets.

On the other hand, objective annotations encompass more definitive diagnostic tests and pathologic assessments of lesions. However, despite their name, they are not free from limitations. For instance, while pathology is the basis of diagnosis, it does not always provide an absolute answer. A pathology result might indicate a very high probability of malignancy and involves a subjective decision on how it is translated into binary malignancy label. This might mean that either the "ground truth" labels are less accurate or a portion of samples is rejected due to low confidence, which, in turn, can introduce a bias relevant for CAD intended use.

Another significant source of annotations is the Electronic Health Records (EHR). These digital records can offer measurements, annotations, or even bounding boxes that are invaluable for certain tasks. However, mining these records, whether done manually or through advanced natural language processing algorithms, is not without challenges. EHR-derived annotations can be noisy and prone to errors, particularly for intricate cases.

Lastly, the interpretation of what constitutes a "true positive" in the context of CAD systems is not obvious. Various methods can yield different results for the same model. For instance, in the domain of lesion detection, the criterion for a true positive could range from measuring distances between centroids of detected objects and the reference, assessing overlap percentages, to determining if a detected object's centroid falls within a reference lesion region. Each methodology can significantly influence system's performance metrics.

The machine learning model development step is undoubtedly required for computer-aided diagnosis system. Given its importance, this stage has gained the most attention within the research community. The increasing number of open-source datasets has provided researchers with easy access to train, test, and fine-tune their models. In addition, the ability to utilize powerful computational resources is no longer restricted to few well-funded labs, institutions or companies. Affordable computing power is accessible to a broader audience, allowing more researchers to experiment, iterate, and innovate. Furthermore, the release of user-friendly ma-



chine learning frameworks has streamlined the process of model development and reduced the learning curve for those new to the field.

Due to the factors above, machine learning models for CAD systems have made great progress, but we need to be careful. It is important to improve models as this helps them be more useful in supporting diagnosis. But we must not forget that this is just one part of the whole process. Sometimes, in the pursuit of making better models, we might forget about how these models will be used or if the source of the data and reference standard are appropriate.

Machine learning model performance evaluation is different from assessing how much a CAD system actually affects doctors in their diagnoses and work. In a CAD system, we are looking at how it assists doctors compared to when they work without it. This assistance might mean, e.g., increasing the sensitivity of detecting lesions or making their work more efficient to allow reading more cases. The way we evaluate CAD systems also depends on how they are intended to be used.

The next level of CAD assessment is the impact on patients. For patients, the main goal is to have a successful treatment. While the earlier evaluations focus on helping doctors diagnose conditions, for a patient, a correct diagnosis is just the first step. Many factors related to the entire healthcare system might affect patient outcomes and it is unclear to what extent a CAD system can help.

Evaluating a CAD system, even in a controlled setting with experts, can be a significant effort. In this context, it is not surprising that most researchers stop at the machine learning model performance alone. To truly understand the CAD system's impact in real-world clinics on both doctors and patients is a large scale project involving many people from different organizations. After all, all necessary precautions must be taken when experimenting with people.

In conclusion, we must remember that CAD systems operate within a much larger healthcare system. There is a long way from a machine learning model that works well on specific data set to a CAD system eventually beneficial to patients.

# Chapter 7

## Overview of academic achievements

The research work presented in this thesis was carried out mainly during the author's appointment at Duke University. He worked for over 2.5 years as an Associate in Research at Carl E. Ravin Advanced Imaging Laboratories, affiliated with the Department of Radiology. He was involved in a number of research projects in the domain of medical imaging and radiology, supervised by Maciej Mazurowski, Ph.D., Associate Professor of Radiology, Computer Science, Electrical and Computer Engineering, and Biostatistics and Bioinformatics at Duke University.

As of November, 2023, Mr. Buda has co-authored eleven articles published in high impact factor journals and three extended abstracts published in Proceedings of Medical Imaging 2020: Computer-Aided Diagnosis Conference. Based on the 2021 Journal Citations Report, the mean impact factor of the journals was 12.7, excluding Radiology: Artificial Intelligence, which was not indexed yet. According to the Web of Science, the total number of citations was over 1600, and the author's h-index was 9. The article published in Neural Networks on the class imbalance problem was the most cited since 2018 for this journal [30]. According to Google Scholar, the total number of citations since 2018 was over 3200, with an h-index of 10 and an i10-index of 10 [31].

One of the most significant achievements in Mr. Buda's research is the publication of the widely-cited article, "A systematic study of the class imbalance problem in convolutional neural networks," in the Neural Networks journal. As of November 2023, this paper has been cited over 2100 times, underlining its influence and the importance of the problem it addresses within the machine learning and computer vision communities. This publication served as the foundational groundwork for subsequent works included in this thesis. This work has significantly contributed to the understanding and development of more robust and reliable convolutional

neural network models for image-based diagnosis in radiology, underscoring Mr. Buda's substantial contribution to the field.

In addition, Mr. Buda was on the organizing committee of DBTex I [32] and DBTex II [33], challenges organized by the International Society for Optics and Photonics, the American Association of Physicists in Medicine, the National Cancer Institute, and the Duke Center for Artificial Intelligence in Radiology [28]. He was a speaker at the "Informatics in Medicine and Biology" seminar, organized by Professor Artur Przelaskowski [34] and at "ML in PL Conference" in Warsaw in 2019 [35]. He also contributed to the academic community as a reviewer for multiple scientific journals, e.g., Neural Networks, Artificial Intelligence Review, Computers in Biology and Medicine, Data Mining and Knowledge Discovery, The Journal of Machine Learning Research, and other.

By making datasets, code, and methodologies publicly available, the author promotes transparency, reproducibility, and collaboration in the field, laying the groundwork for future research directions.

# Bibliography

- [1] Maciej A Mazurowski, **Mateusz Buda**, Ashirbani Saha, and Mustafa R Bashir. Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. *Journal of Magnetic Resonance Imaging*, 49(4):939–954, 2019.
- [2] **Mateusz Buda**, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.
- [3] **Mateusz Buda**, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology & Medicine*, 109:218–225, 2019.
- [4] **Mateusz Buda**, Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep radiogenomics of lower-grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images. *Radiology: Artificial Intelligence*, 2(1), 2020.
- [5] **Mateusz Buda**, Benjamin Wildman-Tobriner, Kerry Castor, Jenny K Hoang, and Maciej A Mazurowski. Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images. *Ultrasound in Medicine & Biology*, 46(2):415–421, 2020.
- [6] **Mateusz Buda**, Benjamin Wildman-Tobriner, Jenny K Hoang, David Thayer, Franklin N Tessler, William D Middleton, and Maciej A Mazurowski. Management of thyroid nodules seen on US images: deep learning may match performance of radiologists. *Radiology*, 292(3):695–701, 2019.
- [7] Benjamin Wildman-Tobriner, **Mateusz Buda**, Jenny K Hoang, William D Middleton, David Thayer, Ryan G Short, Franklin N Tessler, and Maciej A Mazurowski. Using artifi-

- cial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility. *Radiology*, 292(1):112–119, 2019.
- [8] **Mateusz Buda**, Ashirbani Saha, Ruth Walsh, Sujata Ghate, Nianyi Li, Albert Świącicki, Joseph Y Lo, and Maciej A Mazurowski. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. *JAMA network open*, 4(8), 2021.
- [9] Albert Swiecicki, Nicholas Konz, **Mateusz Buda**, and Maciej A Mazurowski. A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis. *Scientific reports*, 11(1):1–13, 2021.
- [10] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.
- [11] Kuniyiko Fukushima. Cognitron: a self-organizing multilayered neural network. *Biological Cybernetics*, 20(3-4):121–136, 1975.
- [12] Kuniyiko Fukushima and Sei Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern recognition*, 15(6):455–469, 1982.
- [13] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.
- [14] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404, 1990.
- [15] Shih-Chung B Lo, Heang-Ping Chan, Jyh-Shyan Lin, Huai Li, Matthew T Freedman, and Seong K Mun. Artificial convolution neural network for medical image pattern recognition. *Neural networks*, 8(7-8):1201–1214, 1995.

- [16] Heang-Ping Chan, Shih-Chung B Lo, Berkman Sahiner, Kwok Leung Lam, and Mark A Helvie. Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network. *Medical physics*, 22(10):1555–1567, 1995.
- [17] 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.
- [18] Lubomir Hadjiiski, Kenny Cha, Heang-Ping Chan, Karen Drukker, Lia Morra, Janne J Näppi, Berkman Sahiner, Hiroyuki Yoshida, Quan Chen, Thomas M Deserno, et al. Aapm task group report 273: Recommendations on best practices for ai and machine learning for computer-aided diagnosis in medical imaging. *Medical Physics*, 50(2):e1–e24, 2023.
- [19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.
- [20] Maciej A Mazurowski, Kal Clark, Nicholas M Czarnek, Parisa Shamsesfandabadi, Katherine B Peters, and Ashirbani Saha. Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data. *Journal of Neuro-oncology*, 133(1):27–35, 2017.
- [21] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, and Roland Wiest. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.
- [22] Maciej A Mazurowski. Radiogenomics: what it is and why it is important. *Journal of the American College of Radiology*, 12(8):862–866, 2015.
- [23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

- [24] Jingxi Weng, Benjamin Wildman-Tobriner, **Mateusz Buda**, Jichen Yang, Lisa M Ho, Brian C Allen, Wendy L Ehieli, Chad M Miller, Jikai Zhang, and Maciej A Mazurowski. Deep learning for classification of thyroid nodules on ultrasound: validation on an independent dataset. *Clinical Imaging*, 99:60–66, 2023.
- [25] Linda Watkins, Greg O’Neill, David Young, and Claire McArthur. Comparison of british thyroid association, american college of radiology tirads and artificial intelligence tirads with histological correlation: diagnostic performance for predicting thyroid malignancy and unnecessary fine needle aspiration rate. *The British journal of radiology*, 94, 2021.
- [26] Lingsze Tan, Ying Sern Tan, and Suzet Tan. Diagnostic accuracy and ability to reduce unnecessary fnac: A comparison between four thyroid imaging reporting data system (tirads) versions. *Clinical Imaging*, 65:133–137, 2020.
- [27] Jenny K Hoang, William D Middleton, Jill E Langer, Kendall Schmidt, Laura B Gillis, Sujith Surendran Nair, Jay A Watts, Randall W Snyder III, Rachita Khot, Upma Rawal, et al. Comparison of thyroid risk categorization systems and fine-needle aspiration recommendations in a multi-institutional thyroid ultrasound registry. *Journal of the American College of Radiology*, 2021.
- [28] Nicholas Konz, **Mateusz Buda**, Hanxue Gu, Ashirbani Saha, Jichen Yang, Jakub Chłędowski, Jungkyu Park, Jan Witowski, Krzysztof J Geras, Yoel Shoshan, et al. A competition, benchmark, code, and data for using artificial intelligence to detect lesions in digital breast tomosynthesis. *JAMA Network Open*, 6(2):e230524–e230524, 2023.
- [29] Jérémie F Cohen, Daniël A Korevaar, Douglas G Altman, David E Bruns, Constantine A Gatsonis, Lotty Hooft, Les Irwig, Deborah Levine, Johannes B Reitsma, Henrica CW De Vet, et al. Stard 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ open*, 6(11):e012799, 2016.
- [30] Elsevier. Most Cited Articles - Neural Networks - Journal. <https://www.journals.elsevier.com/neural-networks/most-cited-articles>. [Online; accessed 30-November-2021].
- [31] Google. Mateusz Buda - Google Scholar. <https://scholar.google.com/citations?user=xJRY-IIAAAAJ>. [Online; accessed 25-November-2023].

- [32] The American Association of Physicists in Medicine. DBTex Challenge. <https://www.aapm.org/GrandChallenge/DBTex/>, . [Online; accessed 25-March-2021].
- [33] The American Association of Physicists in Medicine. DBTex Challenge: Phase 2. <https://www.aapm.org/GrandChallenge/DBTex2/>, . [Online; accessed 25-March-2021].
- [34] Artur Przelaskowski. AIDMED. <https://pages.mim.pw.edu.pl/aidmed/seminaria.html>. [Online; accessed 25-March-2021].
- [35] ML in PL Association. ML in PL 2019 Conference. <https://conference2019.mlinpl.org>. [Online; accessed 30-November-2021].




# **Appendix A**

## **Publications**

### **A.1 Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI**

# Deep Learning in Radiology: An Overview of the Concepts and a Survey of the State of the Art With Focus on MRI

Maciej A. Mazurowski, PhD,<sup>1,2,3\*</sup> Mateusz Buda, MS,<sup>1</sup> Ashirbani Saha, PhD,<sup>1</sup>  and Mustafa R. Bashir, MD<sup>1,4</sup>

Deep learning is a branch of artificial intelligence where networks of simple interconnected units are used to extract patterns from data in order to solve complex problems. Deep-learning algorithms have shown groundbreaking performance in a variety of sophisticated tasks, especially those related to images. They have often matched or exceeded human performance. Since the medical field of radiology mainly relies on extracting useful information from images, it is a very natural application area for deep learning, and research in this area has rapidly grown in recent years. In this article, we discuss the general context of radiology and opportunities for application of deep-learning algorithms. We also introduce basic concepts of deep learning, including convolutional neural networks. Then, we present a survey of the research in deep learning applied to radiology. We organize the studies by the types of specific tasks that they attempt to solve and review a broad range of deep-learning algorithms being utilized. Finally, we briefly discuss opportunities and challenges for incorporating deep learning in the radiology practice of the future.

**Level of Evidence:** 3

**Technical Efficacy:** Stage 1

**J. MAGN. RESON. IMAGING 2019;49:939–954.**

THE FIELD OF DEEP LEARNING encompasses a group of artificial intelligence methods that employ a large number of simple interconnected units to perform complicated tasks. Deep-learning algorithms, rather than using a set of preprogrammed instructions, are capable of learning from large amounts of data. The tasks solved by these algorithms include localizing and classifying objects in images, understanding language, playing games, and many others.<sup>1</sup> While the flagship of deep learning, convolutional neural networks, were first introduced decades ago, only in the last 5 years have astonishing success of these algorithms elevated their status from interesting but impractical ideas to the go-to algorithms in artificial intelligence. In recent years, not only have deep-learning algorithms been able to surpass performance of other methods in artificial intelligence,<sup>2</sup> but in some tasks, such as pneumonia recognition, they have shown performance superior to humans.<sup>3–5</sup>

Arguably, the most well-known achievement of deep learning to date is its performance in the ImageNet competition. ImageNet is a database of more than 14,000,000 annotated natural images containing real-world objects such as cars, animals, and buildings (<http://www.image-net.org>). One of the goals of the competition is to assign each image to one of 1000 predefined categories. When a deep-learning-based algorithm first appeared in the competition in 2012, it dramatically improved the error rate from 0.258 in the previous year (<http://image-net.org/challenges/LSVRC/2011/results>) to 0.153 (<http://image-net.org/challenges/LSVRC/2012/results.html>). The error rate produced by deep-learning-based methods dropped below that achieved by human observers in 2015 for the first time.<sup>5</sup> The performance of deep-learning algorithms for image classification has been improving since then and is now considered comparable to or better than human performance for many tasks.<sup>6–8</sup> Other areas relevant to the topic of this article, where deep-learning algorithms

View this article online at [wileyonlinelibrary.com](http://wileyonlinelibrary.com). DOI: 10.1002/jmri.26534

Received Feb 9, 2018, Accepted for publication Sep 17, 2018.

\*Address reprint requests to: M.A.M., Department of Radiology, Duke University Medical Center, 2424 Erwin Road, Suite 302, Durham, NC 27705.

E-mail: [maciej.mazurowski@duke.edu](mailto:maciej.mazurowski@duke.edu)

From the <sup>1</sup>Department of Radiology, Duke University, Durham, North Carolina, USA; <sup>2</sup>Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina, USA; <sup>3</sup>Duke Medical Physics Program, Duke University, Durham, North Carolina, USA; and <sup>4</sup>Center for Advanced Magnetic Resonance Development, Duke University, Durham, North Carolina, USA

have seen impressive results, include the automatic generation of sophisticated captions for images that consist of full sentences<sup>9</sup> as well as localization and outlining of objects in images.<sup>10,11</sup>

There are likely three reasons for the recent success of deep-learning algorithms: availability of data, increased processing power, and rapid development of algorithms. These are highly connected: availability of large datasets of images and computing power made it possible to demonstrate the strength of the basic concepts of deep learning, and these successes motivated the development of further datasets and algorithms. The availability of graphic processing units (GPUs), which can be used in a multicore model for rapid data processing, has dramatically reduced computation times and enabled larger scientific and technical communities to become involved and to develop even more powerful algorithms, which further advanced the field.

As the primary strength of deep learning has been in image analysis, the potential applications in radiology have become very quickly apparent. The development of algorithms for radiology has shown some inertia due to the time needed for acquisition of the appropriate expertise in the medical imaging community as well as limited availability of large medical imaging datasets. However, the last 2–3 years have seen remarkable productivity in the field. It is now well recognized by both researchers and clinicians that deep learning will play a significant role in radiology.

In this article we begin with a general overview of radiology as the application domain and consider where deep learning could have the most significant impact. Then we introduce the general concepts of deep learning. This is followed by an overview of the recent work in the field, emphasizing developments related to magnetic resonance imaging (MRI). The article closes with remarks regarding the future of deep learning in radiology.

## The Practice of Radiology

Deep-learning techniques (and artificial intelligence algorithms in general) have a tremendous potential to influence the practice of radiology. Unlike most other facets of medicine, nearly all of the primary data utilized in imaging as well as the outputs produced by radiologists (ie, imaging reports) are digital, lending those data to analysis by artificial intelligence algorithms.

One of the most challenging tasks in the interpretation of images is that of disease detection: the rapid differentiation of abnormalities from normal background anatomy. For example, in the interpretation of mammography each radiograph contains thousands of individual focal densities, regional densities, and geometric points and lines that must be interpreted to detect a small number of suspicious or abnormal findings. Fortunately, in order to be useful a

computer algorithm does not have to detect all objects of interest (eg, abnormalities) and be perfectly specific (ie, not mark any normal locations). For example, in screening mammography ~80% of screening mammograms should be read as negative according to the American College of Radiology (ACR) Breast Imaging and Reporting Data System (BI-RADS) guideline. Of the 20% of examinations that trigger additional evaluation, many will ultimately be categorized as negative or benign.<sup>12</sup> An algorithm that could successfully categorize even half of the screening mammograms as definitely negative would dramatically reduce the effort required to interpret a large batch of examinations.

Once an abnormality has been detected, the often-complex task of determining a diagnosis and the disease management implications is undertaken. For focal masses generically, a large number of features must be integrated in order to decide how to appropriately manage the finding. These features can include size, location, signal intensity, borders, heterogeneity, change over time, and others. In some cases, simple criteria have been established and validated for the management of focal findings. For example, most focal lesions in the kidney can be characterized as either simple or minimally complex cysts, which almost uniformly do not require treatment. On the other hand, most lesions in the kidney that are solid are considered to have high malignant potential. Finally, a minority of focal kidney lesions are considered indeterminate and can be managed accordingly. Deep-learning algorithms have the potential to assess a large number of features, including features previously not considered by radiologists, and to arrive at a repeatable conclusion in a fraction of the time required for a human interpreter.

While detection, diagnosis, and characterization of disease receive the primary attention among algorithm developers, another important area where artificial intelligence could contribute is in facilitating the workflow of the radiologists while interpreting images. With the near-complete conversion from printed films to centralized digital Picture Archiving and Viewing Systems (PACS) as well as the availability of multiplanar, multicontrast, and multiphase MRI, radiologists have seen exponential growth in the size and complexity of image data to be analyzed. However, standard PACS systems are not able to reliably organize and present all relevant imaging data to the interpreter for a variety of reasons, including differences in sequence labeling, patient positioning, and anatomy between examinations, variability in modalities used to image the same portion of the anatomy, as well as other factors. In principle, an artificial intelligence algorithm could bring forward sequences from examinations that include the relevant body part(s), detect the image modality and contrast type, and determine the location of the area of interest within the relevant anatomy to reduce the radiologist's effort in performing these relatively mundane tasks.

Finally, computer algorithms might be able to perform medical image interpretation tasks that radiologists do not

perform on a regular basis. For example, the field of radiogenomics<sup>13</sup> aims to find relationships between imaging features of tumors and their genomic characteristics. Examples can be found in breast cancer,<sup>14</sup> glioblastoma,<sup>15</sup> low-grade glioma,<sup>16</sup> and kidney cancer.<sup>17</sup> However, due to its complexity, radiogenomics is not a part of the typical clinical practice of a radiologist. Another example is prediction of outcomes of cancer patients with applications in glioblastoma,<sup>15,18</sup> lower-grade glioma,<sup>16</sup> and breast cancer.<sup>19</sup> While imaging features have a potential to predict patient outcomes, very few are currently used to guide oncological treatment. Deep learning could facilitate the process of incorporating more of the information available from imaging into oncology practice.

## An Introduction to Deep Learning

### Terminology

To understand deep learning, it is helpful to first understand the related concepts of artificial intelligence and machine learning. Artificial intelligence is the most generic of the three terms, comprising a set of computer algorithms that are able to perform complicated tasks or tasks that require intelligence when conducted by humans. Machine learning is a subset of artificial intelligence algorithms which, to perform these complicated tasks, are able to learn from provided data and do not require predefined rules of reasoning. The field of machine learning is very diverse and has already had notable applications in medical imaging.<sup>20</sup> Deep learning is a subdiscipline of machine learning that relies on networks of simple interconnected units. In deep learning models, these units are connected to form multiple layers that are capable of generating increasingly high-level representations of the provided inputs (eg, images). Below, in order to explain the architecture of deep learning models, we introduce the artificial neural network in general and one specific type: the convolutional neural network. Then we detail the process of “learning” as applied to networks, which is the process of incorporating the patterns extracted from data into the deep neural networks.

### Artificial Neural Networks

Artificial neural networks (ANNs) are machine-learning models based on basic concepts dating as far as back as the 1940s, significant development in the 1970s and 1980s, and a period of notable popularity in the 1990s and 2000s, followed by a period of being overshadowed by other machine-learning algorithms. The ANN is based on a concept of an artificial neuron, which is a model of a nerve cell. While many neuron models have been proposed, a typical neuron simply multiplies each input by a certain weight, then adds all the products for all the inputs and applies a simple mathematical function referred to as an activation function, to produce a single output value. An illustration of a neuron and different activation functions is shown in Figs. 1A,B, respectively. An ANN consists of a multitude of interconnected

neurons, usually organized in layers. A simple ANN is illustrated in Fig. 1C. A traditional ANN typically used in the practice of machine learning contains 2 to 3 layers of neurons. Even though each neuron performs a very rudimentary calculation, the interconnected nature of the network allows for the performance of very sophisticated calculations and implementation of very complicated functions.

### Convolutional Neural Networks

Deep neural networks are a special type of ANN. The most common type of deep neural network is a deep convolutional neural network (CNN). Deep CNNs, while inheriting the properties of a generic ANN, also have their own specific features. First, they are “deep,” which is to say that they are typically comprised of 10–30 layers, and in extreme cases could exceed 1000 layers. Second, their neurons are connected such that multiple neurons share weights. This effectively allows the network to perform convolutions (or template matching) of the input image with the filters (defined by the weights) within the CNN. Another special feature of CNNs is that between some layers they perform pooling operations (see Fig. 2), which make the network invariant to small changes in the input data. Finally, CNNs typically use a different nonlinear transformation when generating the output of a neuron as compared with traditional ANNs.

Figure 2 illustrates key concepts for CNNs. Specifically, Fig. 2A demonstrates how a network performs a multiplication of its weights, organized in a matrix by the original pixels within an image. As this multiplication is repeated across different locations in the image, this operation corresponds to filtering of an image where the filters (a.k.a. the convolutional kernels) are defined by the network weights. These layers are referred to as convolutional layers. Figure 2B shows the basic concept of a max pooling layer where a maximum value of multiple neighboring outputs of the previous layer is passed to the next layer. Convolutional layers, pooling layers, and fully connected layers (such as those in the multilayer neural network in Fig. 1C) are the primary components of a CNN. Figure 2C shows an example of a small architecture for a typical CNN. A variety of deep-learning architectures have been proposed, often driven by characteristics of the task at hand (eg, fully convolutional neural networks for image segmentation). Some of these are described in more detail in the section of this article that reviews the current state of the art.

### The Learning Process for CNNs

Above, we described general characteristics of traditional neural networks and deep learning’s flagship, the CNN. Next, we will explore how to make those networks perform useful tasks. This is accomplished in the process referred to as learning or training. The learning process for a CNN simply consists of changing the weights of the individual neurons in response to the provided input data. In the most popular type

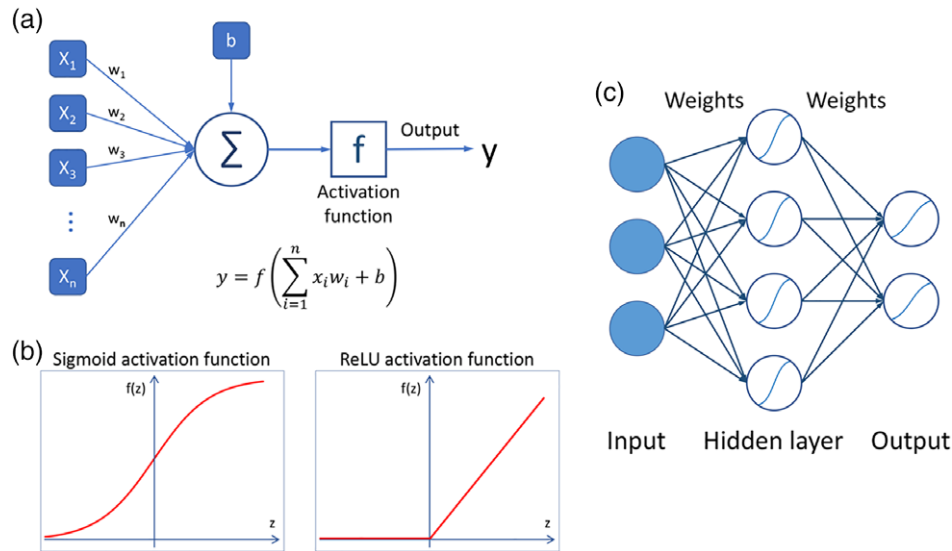


FIGURE 1: A diagram illustrating basic concepts of artificial neural network: (A) a model of a neuron where  $x_1, \dots, x_n$  are the network inputs,  $w_1, \dots, w_n$  are the weights,  $b$  is a bias,  $f$  is the activation function, and  $y$  is the neuron output. (B) Two common activation functions. (C) A model of a simple neural network.

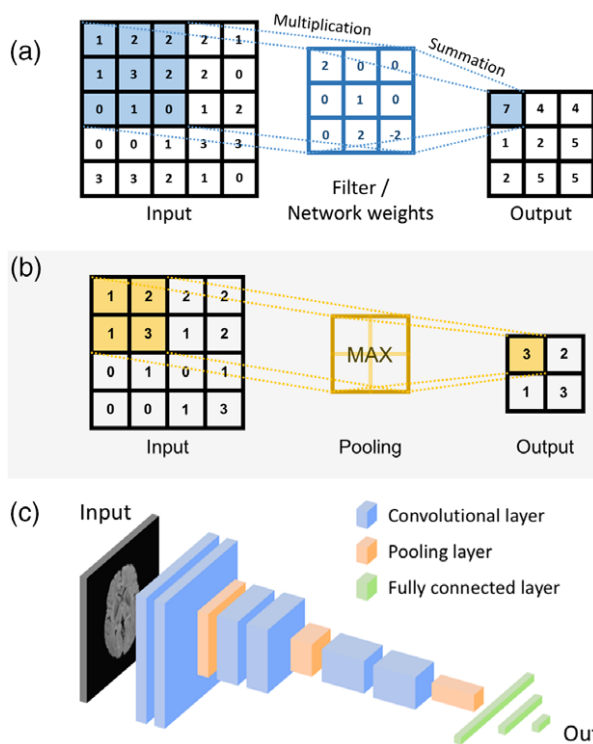
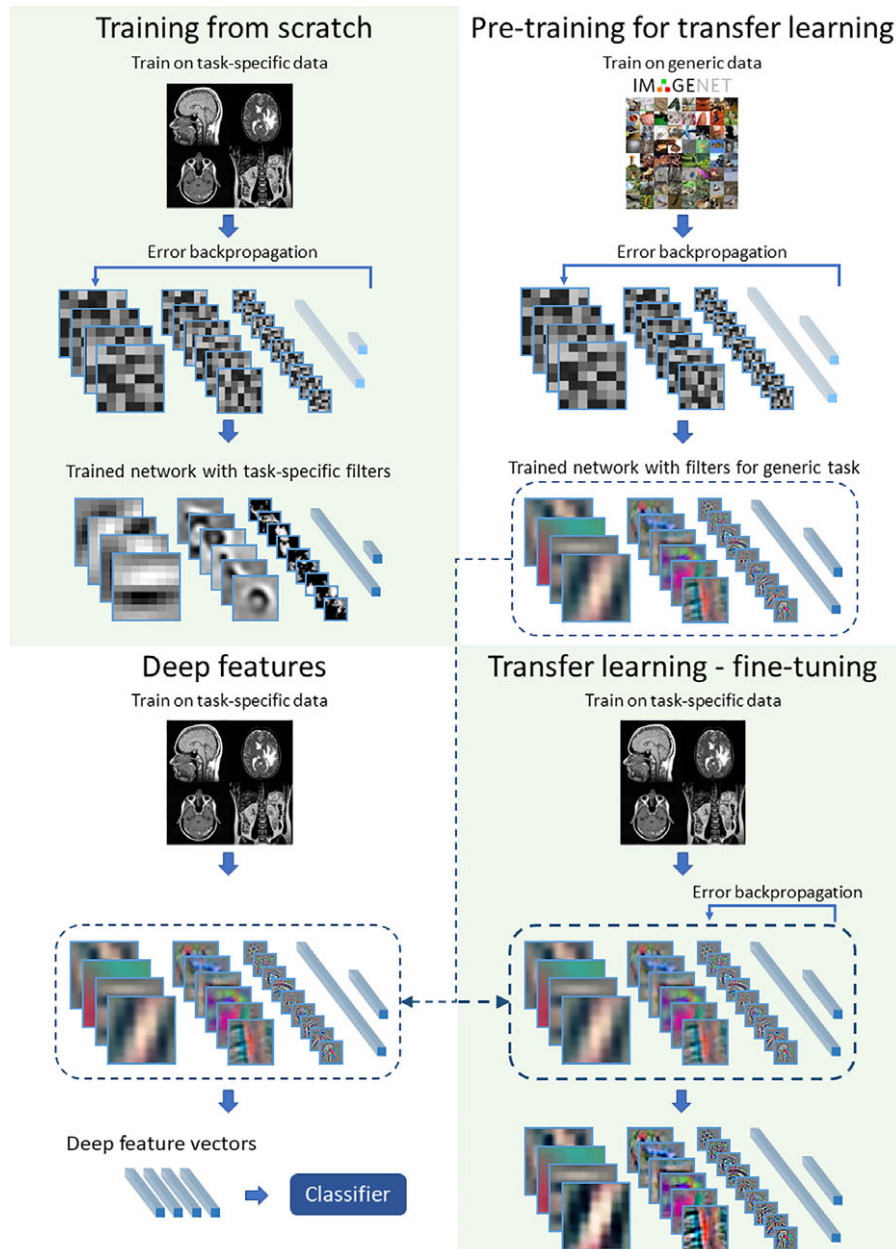


FIGURE 2: A diagram illustrating basic concepts of convolutional neural networks. (a) Convolutional layers: values in the convolutional filters implemented in the network weights (middle column) are multiplied by the pixel values and the products are summed up. (b) A max pooling layer: a maximum pixel value is taken in a given region. (c) An architecture of a simple convolutional neural network including convolutional, pooling, and fully connected layers.

of a learning process, called supervised learning, a training example contains an object of interest (eg, a  $T_2$ -weighted image of a tumor) and a label (eg, the tumor's pathology: benign or malignant). In our example, the image is presented

to the network's input, and the calculation is carried out within the network to produce a predicted summary value (such as a likelihood of malignancy) based on the current weights of the network. Then the network's prediction is compared with the actual label of the object (eg, 0 for benign, 1 for malignant), and an error is calculated. A correction for the error is then propagated through the network to change the values of the network's weights such that the next time the network analyzes this exact example, the error decreases. In practice, the correction of the weights is performed after a group of examples (a batch) are presented to the network. This process is called error backpropagation or stochastic gradient descent. Various modifications of the stochastic gradient descent algorithm have been developed.<sup>21</sup> In principle, this iterative process consists of calculations of error between the output of the model and the desired output and adjusting the weights in the direction where the error decreases.

The most straightforward way of training is to start with a random set of weights and train them using available data specific to the problem being solved (training from scratch). However, given the large number of parameters (weights) in a network, often above 10 million, and a limited amount of training data for a specific task, a network may overtrain (a.k.a. overfit) to the available data (ie, fitting to well to the training set and not generalizing well to test data), resulting in poor performance on test data. Two training methods have been developed to address this issue: transfer learning<sup>22</sup> and off-the-shelf features (a.k.a. deep features).<sup>23</sup> There are many properties of the dataset used for pretraining that affect its usability, eg, similarity of the structures present in the images and the size of the original dataset. However, the quantitative effects of these factors are still a part of ongoing research on transfer learning methods. A diagram comparing training



**FIGURE 3: An illustration of different ways of training in deep neural networks: training from scratch, transfer learning, and deep features**

from scratch with transfer learning and off-the-shelf deep features is shown in Figure 3.

In the transfer learning approach, the network is first trained using a different dataset, for example, an ImageNet collection. Then the network is “fine-tuned” through the addition of training data specific to the problem to be addressed. The idea behind this approach is that performing different visual tasks shares a certain level of processing such as recognition of edges or simple shapes. This approach has been shown successful in, for example, prediction of patient survival time from brain MRI in patients with glioblastoma<sup>24</sup> or in skin lesion classification.<sup>25</sup> Another approach that addresses the issue of limited training data is the deep

“off-the-shelf” features approach that uses CNNs that have been trained on a different dataset to extract features from the images. This is done by using a pretrained network and extracting outputs of layers prior to the network’s final layer. Those layers typically have hundreds or thousands of outputs. Then these outputs are used as inputs to “traditional” classifiers such as linear discriminant analysis, support vector machines, or decision trees. This is similar to transfer learning (and is sometimes considered a part of transfer learning) with the difference being that the final layers of a CNN are replaced by a traditional classifier and the early layers are not additionally trained for the specific task at hand.

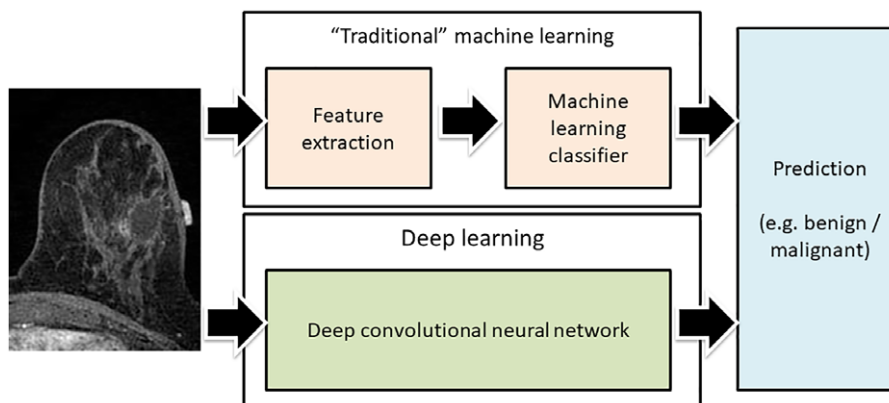


FIGURE 4: An illustration of difference between “traditional” machine learning and deep learning. In the “traditional” machine learning, a set of predefined features is extracted and used by a multivariate classifier. In deep learning the entire image is provided as an input to a neural network, which outputs a decision.

### Deep Learning vs. “Traditional” Machine Learning

Increasingly often we hear a distinction between deep learning and “traditional” machine learning (Fig. 4). The difference is very important, particularly in the context of medical imaging. In traditional machine learning, the first step is typically feature extraction. This means that to classify an object, one must decide which characteristics of an object will be important and implement algorithms that are able to capture these characteristics. A number of sophisticated algorithms in the field of computer vision have been proposed for this

purpose and a variety of size, shape, texture, and other features have been extracted. This process is to a large extent arbitrary, since the machine learning researcher or practitioner often must guess which features will be of use for a particular task and runs the risk of including useless and redundant features and, more important, not including truly useful features. In deep learning, the process of feature extraction and decision making are merged and trainable, and therefore no choices need to be made regarding which features should be extracted; this is decided by the network in the training

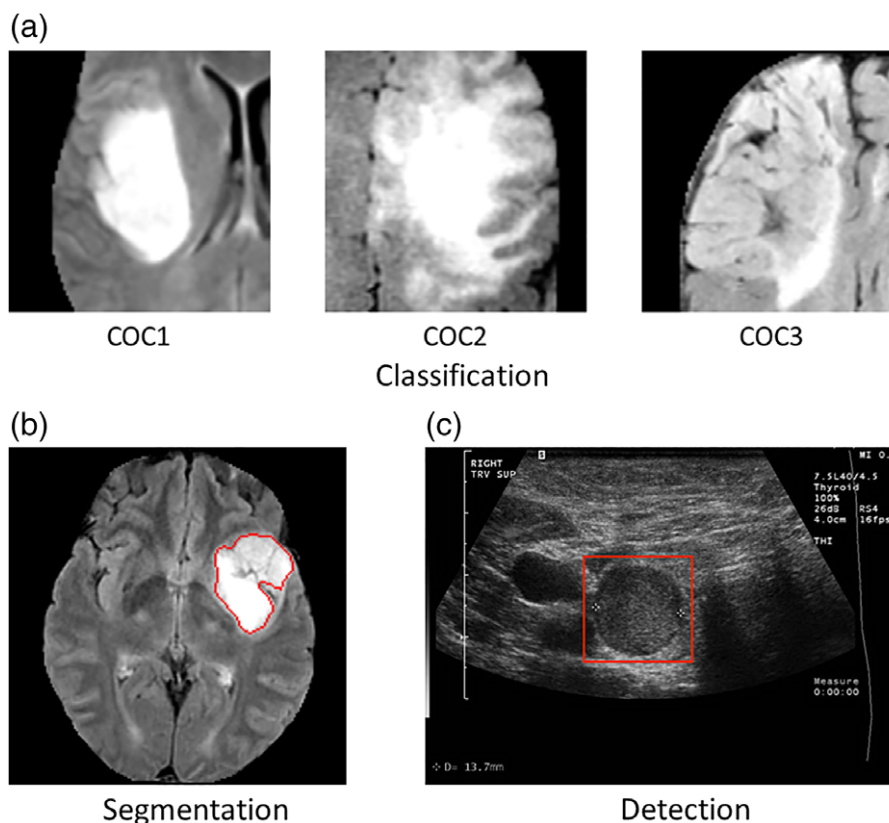


FIGURE 5: Examples of applications of deep neural network to medical images in our laboratory. (A) A classification task in which a CNN was designed to distinguish between different genomic subtypes (cluster of clusters) of lower grade gliomas in MRI. (B) An automatic segmentation of low grade glioma tumors in MRI. (C) A detection of thyroid nodules in ultrasound.



process. However, the cost of allowing the neural network to select its own features is a requirement for much larger training data sets.

## Deep Learning in Radiology: State of the Art

In this section we give an overview of applications of deep learning in radiology. We organized this section by the tasks that the deep-learning algorithms perform. Within each subsection, we describe different methods applied, and, when possible, we systematically discuss the evolution of these methods in recent years. Other recent reviews surveyed the applications of deep learning in broadly understood medical imaging (including pathology)<sup>26</sup> and specifically in brain segmentation in MRI.<sup>27</sup> The summary of the reviewed studies is presented in Table 1.

### Classification

In a classification task, an object is assigned to one of the pre-defined classes. A number of different classification tasks can be found in the domain of radiology, such as: classification of an image or an examination to determine the presence or absence of an abnormality; classification of abnormalities as benign or malignant; classification of cancerous lesions according to their histopathological and genomic features; prognostication; and classification for the purpose of organization radiological data.

Deep learning is becoming the methodology of choice for classifying radiological data. The majority of the available deep-learning classifiers use CNNs with a varying number of convolutional layers followed by fully connected layers. The availability of radiological data is limited as compared with the natural image datasets that have driven the development of deep-learning techniques over the last 5 years. Therefore, many applications of deep learning in medical image classification have resorted to techniques meant to alleviate this issue: off-the-shelf features and transfer learning,<sup>28</sup> discussed in the previous section of this article. Off-the-shelf features have performed well in a variety of domains,<sup>23</sup> and this technique has been successfully applied to medical imaging.<sup>29,30</sup> In Antropova et al.,<sup>29</sup> the authors combined the deep off-the-shelf features extracted from a pretrained deep CNN network with hand-crafted features for determining malignancy of breast lesions in mammography, ultrasound, and MRI and achieved statistically significant improvements in performance compared with existing breast cancer computer-aided diagnosis methods. In Ref. 30, long-term and short-term survival with improved (29%) accuracy was predicted for patients with lung carcinoma by combining off-the-shelf features with the traditional quantitative features. The other strategy, transfer learning, involves fine-tuning of a network pretrained on a different dataset. Transfer learning has been successfully applied to a variety of tasks, such as classification of prostate

MR images to distinguish patients with prostate cancer from patients with benign prostate conditions<sup>31</sup> using MRI. Most of the studies that apply the transfer learning strategy replace and retrain the deepest layer of a network, while the shallow layers are fixed after the initial training. A variant of the transfer learning strategy combines fine-tuning and deep features approaches. It fine-tunes a pretrained network on a new dataset to obtain more task-specific deep feature representations. An ensemble of fine-tuned CNN classifiers was shown to outperform traditional CNNs in predicting the radiological image modality on a test set of 4166 images.<sup>32</sup> A comparison of approaches using deep features and transfer learning with fine-tuning has been shown useful for identifying radiogenomic relationships in breast cancer MRI.<sup>33</sup> Although deep features performed better than transfer learning with the fine-tuning approach, the method faced the issue of training on a small dataset.

When sufficient data are available, an entire deep neural network can be trained from a random initialization (training from scratch). The size of the network to be trained depends on task and dataset characteristics. However, the commonly used architecture in medical imaging is based on AlexNet<sup>2</sup> and VGG<sup>34</sup> with modifications that have fewer layers and weights. Training from scratch has been applied to assessing the presence of Alzheimer's disease based on brain MRI using deep learning.<sup>35</sup> In that study, using the publicly available ADNI cohort, sparse regression models were combined with deep neural networks to achieve higher classification performance compared with several nondeep-learning-based techniques in differentiating Alzheimer's vs. normal controls. Recent advances in the design of CNN architectures have made networks easier to train and more efficient. They have more layers and perform better (in terms of accuracy or area under the curve [AUC]) while having fewer trainable parameters, which reduces the likelihood of overtraining.<sup>36</sup> The most notable examples include Residual Networks (ResNets)<sup>37</sup> and the Inception architecture.<sup>38,39</sup> A shift toward these more powerful networks has also taken place in applications of deep learning to radiology both for transfer learning and training from scratch. Three different ResNets were used to predict methylation of the O6-methylguanine methyltransferase gene status from brain tumor presurgical MRI<sup>40</sup> with an accuracy of 94.9%, which is better than conventional machine learning-based techniques using MRI texture features. In Kim and Mackinnon,<sup>41</sup> the InceptionV3 network was fine-tuned and served as a feature extractor instead of the previously used GoogLeNet to classify wrist radiographs into two categories (with and without fracture). The authors leveraged the data augmentation to generate 11,112 training images from an initial set of 1389 images and obtained an AUC of 0.95 on the test set.

In another approach, auto-encoder<sup>42</sup> or stacked auto-encoder<sup>43</sup> networks were trained from scratch, layer by layer, in an unsupervised way. A stacked denoising auto-



**TABLE 1. Overview of Articles Presented in the Review Split by Task and Organ**

Task	Site	Reference
Classification of breast mass lesions	Breast	Antropova, Huynh, and Giger 2017
Classification of survival groups	Lung	Paul R, Hawkins SH, Balagurunathan Y, Schabath MB, Gillies RJ, Hall LO 2016
Classification of prostate cancer	Prostate	Wang et al. 2017
Classification of image modality	Multiple	Kumar et al. 2017
Classification of genomic subtypes	Breast	Zhu et al. 2017
Classification for Alzheimer's disease	Brain	Suk, Lee, and Shen 2017
Classification of O6-methylguanine methyltransferase gene status	Brain	Korfiatis et al. 2017
Classification of fracture	Wrist	Kim and Mackinnon 2017
Classification for Alzheimer's disease	Brain	Ortiz et al. 2017
Classification for multiple sclerosis	Brain	Yoo et al. 2018
Classification of genomic subtypes	Brain	Akkus, Ali, et al. 2017
Classification of genomic subtypes	Brain	Wachinger, Reuter, and Klein 2017
Segmentation of rectal cancer	Rectum	Trebeschi et al. 2017
Segmentation of brain in fetal US	Fetal brain	Salehi et al. 2017
Segmentation of liver and hepatic lesions	Liver	Christ et al. 2017
Segmentation of liver tumor	Liver	X. Li et al. 2017
Segmentation of prostate gland	Prostate	Clark et al. 2017
Segmentation of sclerosis lesions and gliomas	Brain	McKinley et al. 2016
Segmentation of brain structure	Brain	Mehta and Sivaswamy 2017
Segmentation of prostate	Prostate	Milletari, Navab, and Ahmadi 2016
Segmentation of proximal femur	Proximal femur	Deniz et al. 2017
Segmentation of gliomas	Brain	Shen and Anderson, n.d.
Segmentation of left-ventricle	Brain	Poudel, Lamata, and Montana 2016
Segmentation of pancreas	Pancreas	Cai et al. 2017
Segmentation of left-ventricle	Heart	Avendi, Kheradvar, and Jafarkhani 2016
Detection of sclerosis	Brain	Rey et al. 2002
Detection of lymph nodes	Lymph nodes	Roth et al. 2014
Detection of cerebral microhemorrhage	Brain	Dou et al. 2016
Detection of thoraco-abdominal lymph nodes	Lymph nodes	Shin et al. 2016
Detection of pulmonary embolism	Lung	Tajbakhsh et al. 2016
Detection of masses	Breast	Samala et al. 2016
Detection of intervertebral disc	Spine	Sa et al. 2017
Detection of colitis	Colon	J. Liu et al. 2017
Detection of breast cancer	Breast	Platania et al. 2017

TABLE 1. Continued

Task	Site	Reference
Detection of breast tumor	Breast	Cao et al. 2017
Detection of pulmonary lung nodules	Lung	N. Li et al. 2017
Detection of 3D anatomy in chest	Chest	de Vos et al. 2016
Detection of knee cartilage	Knee	Prasoon et al. 2013
Detection of sclerotic metastases	Brain	Roth, Lu, et al. 2016
Detection of lymph nodes	Lymph nodes	Roth, Lu, et al. 2016
Detection of colonic polyp	Colon	Roth, Lu, et al. 2016
Detection of fractures on spine	Spine	Roth, Wang, et al. 2016
Registration of T1-T2 MRI of the neonatal brain	Brain	Simonovsky et al. 2016
Registration of brain MRI	Brain	Maes et al. 1997
Correction of respiratory motion	Abdomen	Lv et al. 2017
Registration of cardiac cine MRI	Heart	de Vos et al. 2017
Reconstruction of 7T-like MRI	Brain	Bahrami et al. 2016
Reconstruction of MRI	Heart	Schlemper et al. 2017
Reconstruction of compressed sensed MRI	Abdomen	Yang et al. 2017
Synthesis of MRI	Brain	Chartsias et al. 2017
Reconstruction of CT images from low-dose CT images	Multiple	H. Chen et al. 2017
Reconstruction of CT images from low-dose CT images	Abdomen	Kang, Min, and Ye 2017
Generation of CT images from MRI	Pelvis	Nie et al. 2016
Generation of CT images from MRI	Brain	Han 2017
Prediction of PET pattern from MRI	Brain	R. Li et al. 2014
Super-resolution in MRI	Heart	Oktay et al. 2016
Enhancement of DCE-MRI	Brain	Benou et al. 2017
Denoising of 3D MRI	Brain	Jiang et al. 2017
Content-based image retrieval	Multiple	Qayyum et al. 2017
Automatic objective image quality assessment	Fetal	Wu Lingyun, Cheng Jie-Zhi, Li Shengli, Lei Baiying, Wang Tianfu 2017
Automatic objective image quality assessment	Heart	Abdi AH, Luong C, Tsang T, Allan G, Nouranian S, Jue J, Hawley D, Fleming S, Gin K, Swift J 2017
Diagnostic quality assessment of MRI	Liver	Esses et al. 2017

encoder with backpropagation was used in Ortiz et al.<sup>44</sup> to determine the presence of Alzheimer's disease. Auto-encoders and stacked auto-encoders can also be used to extract feature representations (similar to the deep features

approach) from hidden layers for further classification. Such feature representation has also been used in the identification of multiple sclerosis lesions in using MRI and myelin maps jointly.<sup>45</sup>

## Segmentation

In an image segmentation task, an image is divided into different regions in order to separate distinct parts or objects. In MRI, the common applications are segmentation of organs, substructures, or lesions, often as a preprocessing step for feature extraction and classification.<sup>46,47</sup> Below, we discuss different types of deep learning approaches used in segmentation tasks in a variety of radiological images.

The most straightforward and still widely used method for image segmentation is classification of individual voxels based on small image patches (both 2D and 3D patches) extracted around the classified voxel. This approach has found use in various segmentation problems; for example, brain tumor segmentation,<sup>48–50</sup> white matter segmentation in multiple sclerosis patients,<sup>51</sup> segmentation of normal components of brain anatomy,<sup>52</sup> and rectal cancer segmentation.<sup>53</sup> It allows for using the same network architectures and solutions that are known to work well for classification; however, there are some shortcomings to this method. The primary issue is that these methods are computationally inefficient, since they process overlapping parts of images multiple times. Another drawback is that each voxel is segmented based on a limited-size context window and ignores the wider context. In some cases, some global information, eg, pixel location or relative position to other image parts, may be needed to correctly assign its label.

One approach that addresses the shortcomings of the voxel-based segmentation is a fully convolutional neural network (fCNN).<sup>54</sup> Networks of this type process the entire image (or large portions of it) at the same time and output a 2D map of labels (ie, a segmentation map) instead of a label for a single pixel. A very important advantage of fCNNs over the voxel-based approach is avoiding many repeated convolutions by analyzing a large portion of the image and providing the segmentation label for all the voxels at the same time. Example architectures that were successfully used in both natural images and radiology applications are encoder–decoder architectures such as U-Net<sup>55–57</sup> or Fully Convolutional DenseNet.<sup>58–60</sup> Various adjustments to these types of architectures have been developed that mainly focus on connections between the encoder and decoder parts of the networks, called skip connections. An fCNN was applied in Clark et al<sup>61</sup> in radiology that included prostate gland segmentation in diffusion-weighted MRI. Although a relatively small dataset of over 100 cases was used, the segmentation quality as evaluated with a Dice similarity coefficient was 0.89. In another study,<sup>62</sup> an fCNN was used for segmentation of multiple sclerosis lesions and gliomas in MRI slices of axial, coronal, and sagittal planes separately. In addition to differences in building blocks for fCNNs, different optimization functions have been explored that account for class imbalance (remarkable differences among the number of examples in each class), which is common in medical datasets.<sup>63</sup> In Mehta

and Sivaswamy,<sup>64</sup> weighted cross-entropy loss was used for brain structure segmentation in MRI. The proposed method did not require any postprocessing and offered on average 10 times faster processing of large MRI volumes compared with other tested methods.

In order to segment 3D data, it is common to process data as 2D slices and then combine the 2D segmentation maps into a 3D map, since 3D fCNNs are significantly larger in terms of trainable parameters and as a result require significantly larger amounts of data. Nevertheless, these obstacles can be overcome, and there are successful applications of 3D fCNNs in radiology, eg, V-Net for prostate segmentation from MRI,<sup>65</sup> 3D U-Net<sup>66</sup> for segmentation of the proximal femur for assessing osteoporosis,<sup>67</sup> and brain glioma segmentation.<sup>68</sup>

Finally, a deep learning approach that has found some application in medical imaging segmentation is recurrent neural networks. In Poudel et al,<sup>69</sup> the authors applied a recurrent fCNN for left-ventricle segmentation in multislice cardiac MRI to leverage interslice spatial dependences. Similarly, Cai et al<sup>70</sup> used a long short / term memory (LSTM)<sup>71</sup> type of recurrent neural network trained end-to-end together with fCNN to take advantage of 3D contextual information for pancreas segmentation in MR images. In addition, they proposed a novel loss function that directly optimizes a widely used segmentation metric, the Jaccard Index.<sup>72</sup>

## Detection

Detection is a task of localizing and pointing out (eg, using a rectangular box) an object in an image. In radiology, detection is often an important step in the diagnostic process that identifies an abnormality (such as a mass or a nodule), an organ, an anatomical structure, or a region of interest for further classification or segmentation.<sup>73,74</sup> Here we discuss the common architectures used for various detection tasks in radiology along with example specific applications.

The most common approach to detection for 2D data is a two-phase process that requires training of two models. The first phase identifies all suspicious regions that may contain the object of interest. The requirement for this phase is high sensitivity,<sup>75</sup> and therefore it usually produces many false positives. A typical deep-learning approach for this phase is a regression network for bounding box coordinates based on architectures used for classification.<sup>76,77</sup> The second phase is simply classification of subimages extracted in the previous step. In some applications, only one of the two steps uses deep learning. This strategy has been applied in cerebral microhemorrhage detection using a large dataset of 320 MRI volumes and achieved 93% sensitivity.<sup>78</sup>

The classification step, when utilizing deep learning, is often performed using transfer learning. The models are often pretrained on natural images, for example for thoraco-abdominal lymph node detection in Shin et al<sup>79</sup> and

pulmonary embolism detection in computed tomography (CT) pulmonary angiogram images.<sup>28</sup> In other applications, models have been pretrained using other medical imaging datasets to detect masses in digital breast tomosynthesis images.<sup>80</sup> The same network architectures can be used for the second phase, as in a regular classification task (eg, VGG,<sup>34</sup> GoogLeNet,<sup>81</sup> Inception,<sup>38</sup> ResNet<sup>37</sup>, depending on the needs of a particular application.

While in the two-phase detection process the models are trained separately for each phase, in the end-to-end approach one model encompassing both phases is trained. An end-to-end architecture that has proved to be successful in object detection in natural images, and was recently applied to medical imaging, is the Faster Region-based Convolutional Neural Network.<sup>10</sup> It uses a CNN to obtain a feature map that is shared between a region proposal network that outputs bounding box candidates, and a classification network that predicts the category of each candidate. It was recently applied for intervertebral disc detection in X-ray images<sup>82</sup> and detection of colitis on CT images.<sup>83</sup>

Another approach to detection is a single-phase detector that eliminates the first phase of region proposals. Examples of popular methods that were first developed for detection in natural images and rely on this approach are You Only Look Once (YOLO),<sup>84</sup> Single Shot MultiBox Detector,<sup>85</sup> and RetinaNet.<sup>11</sup> In the context of radiology, a YOLO-based network called BC-DROID has been developed for region of interest detection in breast mammograms.<sup>86</sup> Single Shot MultiBox Detector has been employed for breast tumor detection in ultrasound images, outperforming other evaluated deep-learning methods that were available at the time.<sup>87</sup> Li et al<sup>88</sup> applied the same network for detection of pulmonary lung nodules in CT images. The above-mentioned methods and architectures were widely adapted for natural images and some medical imaging modalities, eg, CT, mammograms, X-rays, however, are still uncommonly applied in object detection using MRI.

In the examples above, 2D data have typically been used. For 3D imaging volumes, which are most commonly encountered in CT and MRI, the results obtained from 2D processing can be combined to produce the final 3D bounding box. As an example, in de Vos et al<sup>89</sup> the authors performed detection of 3D anatomy in chest CT images by processing data slice-by-slice in one direction. Combining output from different planes was performed in several studies. Most of the them<sup>90–92</sup> used orthogonal planes of MRI and CT images performing detection in each direction separately. The results can then be combined in different ways, eg, by an algorithm based on output probabilities<sup>89</sup> or using another machine-learning method like random forest.<sup>88</sup> An alternative method for 3D detection has been proposed for automatic detection of lymph nodes by concatenating coronal, sagittal, and axial views as a single three-channel image.<sup>75</sup>

## Other Tasks in Radiology

While the majority of the applications of deep learning in radiology have been in classification, segmentation, and detection, other medical imaging-related problems have found some solutions in deep learning. Due the variety of those problems, there is no unifying methodological framework for these solutions. Therefore, below we organize the examples according to the problem that they attempt to address.

**IMAGE REGISTRATION.** In this task two or more images (often 3D volumes), typically of different types (eg, T<sub>1</sub>-weighted and T<sub>2</sub>-weighted image sets) must be spatially aligned such that the same location in each image represents the same physical location in the depicted organ. Several approaches can be used to address the problem. In one approach, similarity measures between image patches taken from the images of interest are calculated and used to register the image sets. Simonovsky et al<sup>93</sup> used deep learning to learn a similarity measure from T<sub>1</sub>-T<sub>2</sub> MR image pairs of the adult brain and tested it to register T<sub>1</sub>-T<sub>2</sub> MRI interpatient images of the neonatal brain. This similarity measure performed better than the standard measure, called mutual information, which is widely used in registration.<sup>94</sup> In another deep-learning-based approach to image registration, the deformation parameters between image pairs are directly learned using misaligned image pairs. A CNN-based network was trained to correct respiratory motion in 3D abdominal MR images<sup>95</sup> by predicting spatial transforms. All of these techniques are supervised regression techniques, as they were trained using ground truth deformation information. In another approach,<sup>96</sup> which was unsupervised, a CNN was trained end-to-end to generate a spatial transformation that minimized dissimilarity between misaligned image pairs.

**IMAGE GENERATION/RECONSTRUCTION.** Acquisition and hardware parameters can strongly affect the visual quality and detail of images obtained using the same modality. First, we discuss the applications that synthesize images generated using different acquisition parameters within the same modality. In Bahrami et al,<sup>97</sup> 7T-like images were generated from 3T MR images by training a CNN with patches centered around voxels in the 3T MR images. Undersampled (in *k*-space) cardiac MRIs were reconstructed using a deep cascade of CNNs in Schlemper et al.<sup>98</sup> A real-time method to reconstruct compressed sensed MRI using generative adversarial networks (GAN) has also been proposed.<sup>99</sup> In another approach,<sup>100</sup> in order to synthesize brain MRIs based on other MRI sequences in the same patient, convolutional encoders were built to generate a latent representation of images. Then, based on that representation, a sequence of interest was generated. Reconstruction of “normal-dose” CT images from low-dose CT images (which are degraded in comparison to normal-dose images) has been performed using

patch-by-patch mapping of low-dose images to high-dose images using a shallow CNN.<sup>101</sup> In contrast, a deep CNN has been trained with low-dose abdominal CT images for reconstruction of normal-dose CT.<sup>102</sup>

Deep learning has also been applied to synthesizing images of different modalities. For example, CT images have been generated using MR images by adopting an fCNN to learn an end-to-end nonlinear mapping between pelvic CT and MR images.<sup>103</sup> Synthetic CT images of the brain have also been generated from a single T<sub>1</sub>-weighted MR image set.<sup>104</sup> In another application to aid a classification framework for Alzheimer's disease diagnosis with missing positron emission tomography (PET) scans, PET patterns were predicted from MRI using CNN.<sup>105</sup>

**IMAGE ENHANCEMENT.** Image enhancement aims to improve different characteristics of the image such as resolution, signal-to-noise-ratio, and necessary anatomical structures (by suppressing unnecessary information) through various approaches such as superresolution and denoising.

Superresolution of images has particularly been explored in cardiac and lung imaging. Three-dimensional near-isotropic cardiac and lung images often require long scan times in comparison to the time the subject can hold his or her breath. Thus, multiple thick 2D slices are acquired instead and the superresolution methodology is applied to improve the through-plane resolution of the images. A deep cascade of CNNs has been shown to preserve anatomical structure up to 11-fold undersampling using cardiac MRI.<sup>106</sup> In another study<sup>107</sup> using CT, a single image superresolution approach based on CNN was applied in a publicly available chest CT image dataset to generate high-resolution CT images, which are preferred for interstitial lung disease detection. This method outperformed the traditional compressed sensing-based approaches used in MR image reconstruction. In another study,<sup>108</sup> to synthesize thin slice knee MRIs from thick slice knee MRIs, the proposed CNN-based approach showed improved qualitative and quantitative performance over the state-of-the-art techniques in a test set of 17 patients.

Image enhancement through denoising application of using deep learning has also been described in Benou et al,<sup>109</sup> where the authors performed denoising of DCE-MRI images of a brain (for stroke and brain tumors) by training an ensemble of deep auto-encoders using synthesized data. Removal of Rician noise in real and synthetic 3D MR images using a deep CNN aided with residual learning can be performed by excluding the traditional steps of optimization and estimation of the noise level parameter.<sup>110</sup> An encoder-decoder CNN architecture<sup>111</sup> was used to denoise the noisy uptake signal between a precontrast MR sequence (zero gadolinium dose for contrast) and a 10% low-dose postcontrast MR sequence of brain. With the help of this model, full contrast high-quality

postcontrast sequences were reconstructed from sequences with 10-fold reduction in contrast dose for different pathologies (including glioma) in brain for 50 patients.

**CONTENT-BASED IMAGE RETRIEVAL.** In the most typical version of this task, the algorithm, given a query image, finds the most similar images in a given database. To accomplish this task, a deep CNN can first be trained to distinguish between different organs.<sup>112</sup> Then features from the three fully connected layers in the network are extracted for the images in the set from which the images were retrieved (evaluation dataset). The same features can then be extracted from the query image and compared with those of the evaluation dataset to retrieve the image.

**OBJECTIVE IMAGE QUALITY ASSESSMENT.** Objective quality assessment measures of medical images aim to classify an image to be of satisfactory or unsatisfactory quality for subsequent tasks. Objective quality measures of medical images diagnosis aid in better treatment.<sup>113</sup> Image quality of fetal ultrasound has recently been predicted using CNN.<sup>114</sup> Another study attempted to reduce the data acquisition variability in echocardiograms using a CNN trained on the quality scores assigned by an expert radiologist.<sup>115</sup> As another example, simple CNN architecture has been reported for classifying T<sub>2</sub>-weighted liver MR images as diagnostic or non-diagnostic quality by CNN.<sup>116</sup>

## Main Challenges and Pitfalls in Development of Deep-Learning Algorithms

While applications of deep learning in medical imaging show tremendous promise, there are some challenges and potential pitfalls, and caution should be exercised in research on the topic. One of the principal challenges is the availability of data. While millions of training examples are available for problems related to natural images, the datasets for medical images are typically much smaller, with a typical number of patients in the hundreds range. This, combined with the large number of parameters in a deep neural network that require optimization, results in a high risk of overtraining and subsequent low performance on data that were not used in the training process. Some solutions that can help alleviate this issue are pretraining of the models with other datasets, use of smaller models, and augmentation of the data by including slight alterations of the original images in the dataset. A related issue is often a very small number of cases with a disease (eg, cancer) as compared with healthy patients. This issue, referred to as class imbalance, can lead to highly diminished performance.<sup>118</sup> Some solutions have been proposed for this issue such as a higher rate of sampling of the examples from the minority class for training.<sup>118</sup> However, despite the solutions that have been proposed to address small dataset size

and class imbalance, these remain important challenges to the use of deep learning in radiology.

Given the high risk of overtraining, there is a high likelihood of reporting performance that does not reflect the true ability of a model to classify/predict/segment when the validation is not conducted properly. Although evaluation of models through splitting datasets into a training dataset, which is used for the development of the model and a test set used for estimating the model's performance, as well as cross-validation (splitting the dataset into training and test sets and combining the evaluation results) can provide a fairly accurate estimate of generalization performance, these methods also have limitations. Therefore, sharing the developed models for testing by other institutions can facilitate further development by testing reproducibility and can increase the confidence of the scientific community in these models.

Finally, when aiming to develop deep-learning models that could be used clinically, one must ensure a proper validation setup in the experimental validation of the models. This is highly challenging and sometimes overlooked. It requires not only posing a clinical question that is of significance and answers that change clinical decision-making, but also careful curation of the dataset to include only those patients who are relevant to the question and precise definitions of other nonimaging variables such as pathology/genomic markers and patient outcomes. This requires a close and continuous collaboration with clinicians and/or other experts in a given application field at many stages of the development, as well as a strong understanding of the clinical reality of the problem by the technical expert. While those and other issues pose challenges in the development of deep-learning models, none of them are insurmountable.

## Future of Deep Learning in Radiology

There is a general agreement that deep learning will play a role in the future practice of radiology, and MRI specifically. Some predict that deep-learning algorithms will conduct mundane tasks, leaving radiologists with more time to focus on intellectually demanding challenges. Others believe that radiologists and deep-learning algorithms will work hand-in-hand to deliver performance superior to either alone. Finally, some predict that deep-learning algorithms will replace radiologists (at least in their image interpretation capacity) altogether.

Incorporation of deep learning in radiology will be associated with multiple challenges. First, and currently foremost, is the technological challenge. While deep learning has shown extraordinary promise in other image-related tasks, the results in radiology are still far from showing that deep-learning algorithms will replace a radiologist in the entire scope of their diagnostic work. Some recent studies suggest performance of these algorithms comparable to expert humans in narrowly

defined tasks, but these results are only applicable to a very small minority of the tasks that radiologists perform.<sup>4,119–125</sup> This is likely to change in the upcoming years, given the rapid progress in implementing the deep-learning algorithms in the realm of radiology.

Implementation of deep learning in radiology practice also poses legal and ethical challenges. Primarily: Who will be responsible for the mistakes that a computer will make? While this is a difficult question, similar questions have been posed and resolved when other technologies were introduced; for example, elevators and cars. Since artificial intelligence penetrates various areas of human activity, questions of this type will likely be studied and answers proposed in the coming years.

Other challenges will include patient acceptance or non-acceptance of a human's not being involved in the process of interpreting their images (regardless of the performance) as well as regulatory issues. Finally, an important practical issue is how to incorporate deep-learning algorithms into the radiology workflow in order to improve, rather than disrupt, the radiology practice.

## Conclusion

In summary, in this article we have discussed the principles of deep learning as well as the current practice of radiology to elucidate how these new algorithms may be incorporated into the radiologists' workflow. We have discussed the progress and state of art in the field. Finally, we have discussed some challenges and questions related to implementation of deep learning in the current practice of imaging. All signs show that deep learning will play a significant role in radiology. The next 5 years will be a very exciting time in the field that may see many questions stated in this article answered through a collaboration of machine learning scientists and radiologists.

## Acknowledgment

Contract grant sponsor: National Institutes of Biomedical Imaging and Bioengineering; Contract grant number: 5 R01 EB021360.

The authors thank Gemini Janis for reviewing and editing the article.

## References

1. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. *Neurocomputing* 2017; 234:11–26.
2. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Adv Neural Inf Process Syst* 2012; 1097–1105.
3. Dodge S, Karam L. A study and comparison of human and deep learning recognition performance under visual distortions. *arXiv Prepr arXiv:170502498* 2017.

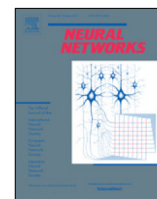
4. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv Prepr arXiv171105225* 2017.
5. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proc IEEE Int Conf Comput Vis* 2015;1026–1034.
6. Taigman Y, Yang M, Ranzato M, Wolf L. Deepface: Closing the gap to human-level performance in face verification. In: *Proc IEEE Conf Comput Vis Pattern Recognit* 2014;1701–1708.
7. Wu R, Yan S, Shan Y, Dang Q, Sun G. Deep image: Scaling up image recognition. *arXiv Prepr arXiv150102876* 2015.
8. Chung JS, Senior AW, Vinyals O, Zisserman A. Lip reading sentences in the wild. In: *CVPR* 2017;3444–3453.
9. Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In: *Proc IEEE Conf Comput Vis Pattern Recognit* 2015;3128–3137.
10. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: *Adv Neural Inf Process Syst* 2015;91–99.
11. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. *arXiv Prepr arXiv170802002* 2017.
12. Ghatge SV, Soo MS, Baker JA, Walsh R, Gimenez EI, Rosen EL. Comparison of recall and cancer detection rates for immediate versus batch interpretation of screening mammograms. *Radiology* 2005;235:31–35.
13. Mazurowski MA. Radiogenomics: What it is and why it is important. *J Am Coll Radiol* 2015;12:862–866.
14. Mazurowski MA, Zhang J, Grimm LJ, Yoon SC, Silber JL. Radiogenomic analysis of breast cancer: Luminal B molecular subtype is associated with enhancement dynamics at MR imaging. *Radiology* 2014;273:365–372.
15. Gutman DA, Cooper LA, Hwang SN, et al. MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. *Radiology* 2013;267:560–569.
16. Mazurowski MA, Clark K, Czarnek NM, Shamsesfandabadi P, Peters KB, Saha A. Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with The Cancer Genome Atlas data. *J Neurooncol* 2017;1–9.
17. Karlo CA, Di Paolo PL, Chaim J, et al. Radiogenomics of clear cell renal cell carcinoma: associations between CT imaging features and mutations. *Radiology* 2014;270:464–471.
18. Mazurowski MA, Desjardins A, Malof JM. Imaging descriptors improve the predictive power of survival models for glioblastoma patients. *Neuro Oncol* 2013;15:1389–1394.
19. Mazurowski MA, Grimm LJ, Zhang J, et al. Recurrence-free survival in breast cancer is associated with MRI tumor enhancement dynamics quantified using computer algorithms. *Eur J Radiol* 2015;84:2117–2122.
20. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. *RadioGraphics* 2017;37:505–515.
21. Ruder S. An overview of gradient descent optimization algorithms. *arXiv:1609.04747* 2016.
22. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: *Adv Neural Inf Process Syst* 2014;3320–3328.
23. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: an astounding baseline for recognition. In: *Proc IEEE Conf Comput Vis Pattern Recognit Work* 2014;806–813.
24. Ahmed KB, Hall LO, Goldgof DB, Liu R, Gatenby RA. Fine-tuning convolutional deep features for MRI based brain tumor classification. In: *Med Imaging 2017 Comput Diagnosis* 10134; 2017:101342E.
25. Esteva A, Kuprel B, Novoa R, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:686.
26. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60–88.
27. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. *J Digit Imaging* 2017;30:449–459.
28. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans Med Imaging* 2016;35:1299–1312.
29. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med Phys* 2017;44:5162–5171.
30. Paul R, Hawkins SH, Balagurunathan Y, et al. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography* 2016;2:388–395.
31. Wang X, Yang W, Weinreb J, et al. Searching for prostate cancer by fully automated magnetic resonance imaging classification: Deep learning versus non-deep learning. *Sci Rep* 2017;7:15415.
32. Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE J Biomed Heal Informatics* 2017;21:31–40.
33. Zhu Z, Albadawy E, Saha A, Zhang J, Harowicz MR, Mazurowski MA. Deep learning for identifying radiogenomic associations in breast cancer. *arXiv Prepr arXiv171111097* 2017.
34. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv Prepr arXiv14091556* 2014.
35. Suk H-I, Lee S-W, Shen D. Deep ensemble learning of sparse regression models for brain disease diagnosis. *Med Image Anal* 2017;37:101–113.
36. Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications. *arXiv Prepr arXiv160507678* 2016.
37. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proc IEEE Conf Comput Vis Pattern Recognit* 2016;770–778.
38. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In: *AAAI* 2017;4278–4284.
39. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *Proc IEEE Conf Comput Vis Pattern Recognit* 2016;2818–2826.
40. Korfiatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ. Residual deep convolutional neural network predicts MGMT methylation status. *J Digit Imaging* 2017;30:622–628.
41. Kim DH, Mackinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol* 2018;73:439–445.
42. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006;313:504–507.
43. Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. In: *Adv Neural Inf Process Syst* 2007;153–160.
44. Ortiz A, Munilla J, Martínez-Murcia FJ, Górriz JM, Ramírez J. Learning longitudinal MRI patterns by SICE and deep learning: Assessing the Alzheimer's disease progression. In: *Commun Comput Inf Sci* 2017;723:413–424.
45. Yoo Y, Tang LYW, Brosch T, et al. Deep learning of joint myelin and T1w MRI features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls. *NeuroImage Clin* 2018;17:169–178.
46. Akkus Z, Ali I, Sedl   J, et al. Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence. *J Digit Imaging* 2017;30:469–476.
47. Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Sci Rep* 2017;7:5467.
48. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with deep neural networks. *Med Image Anal* 2017;35:18–31.
49. Milletari F, Ahmadi S-A, Kroll C, et al. Deep learning for segmentation of deep brain regions in MRI and ultrasound. *Comput Vis Image Underst* 2017 [Epub ahead of print].

50. Hussain S, Anwar SM, Majid M. Brain tumor segmentation using cascaded deep convolutional neural network. In: Eng Med Biol Soc (EMBC), 2017 39th Annu Int Conf IEEE 2017;1998–2001.
51. Valverde S, Cabezas M, Roura E, et al. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. Neuroimage 2017;155:159–168.
52. Wachinger C, Reuter M, Klein T. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. Neuroimage 2018;170:434–445.
53. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. Sci Rep 2017;7:5301.
54. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proc IEEE Conf Comput Vis Pattern Recognit 2015;3431–3440.
55. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Int Conf Med Image Comput Comput Interv 2015;234–241.
56. Salehi SSM, Hashemi SR, Velasco-Annis C, et al. Real-time automatic fetal brain extraction in fetal MRI by deep learning. arXiv Prepr arXiv171009338 2017.
57. Christ PF, Ettlinger F, Grün F, et al. Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. arXiv Prepr arXiv170205970 2017.
58. Jégou S, Drozdal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Comput Vis Pattern Recognit Work (CVPRW), 2017 IEEE Conf 2017;1175–1183.
59. Li X, Chen H, Qi X, Dou Q, Fu C-W, Heng PA. H-DenseUNet: Hybrid densely connected UNet for liver and liver tumor segmentation from CT volumes. arXiv Prepr arXiv170907330 2017.
60. Chen L, Wu Y, DSouza AM, Abidin AZ, Xu C, Wismüller A. MRI tumor segmentation with densely connected 3D CNN. arXiv Prepr arXiv1802.02427 2018.
61. Clark T, Wong A, Haider MA, Khalvati F. Fully deep convolutional neural networks for segmentation of the prostate gland in diffusion-weighted MR images. In: Int Conf Image Anal Recognit 2017;97–104.
62. McKinley R, Weper R, Gundersen T, et al. Nabla-net: A deep Dag-like convolutional architecture for biomedical image segmentation. In: Int Work Brainlesion Glioma, Mult Sclerosis, Stroke Trauma Brain Inj 2016; 119–128.
63. Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learn Med Image Anal Multimodal Learn Clin Decis Support; Springer. 2017:240–248.
64. Mehta R, Sivaswamy J. M-net: A convolutional neural network for deep brain structure segmentation. In: Biomed Imaging (ISBI 2017), 2017 IEEE 14th Int Symp 2017;437–440.
65. Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vis (3DV), 2016 Fourth Int Conf 2016;565–571.
66. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Int Conf Med Image Comput Comput Interv 2016;424–432.
67. Deniz CM, Hallyburton S, Welbeck A, Honig S, Cho K, Chang G. Segmentation of the proximal femur from MR images using deep convolutional neural networks. arXiv Prepr arXiv170406176 2017.
68. Shen L, Anderson T. Multimodal brain MRI tumor segmentation via convolutional neural networks. <https://www.semanticscholar.org/paper/Multimodal-Brain-MRI-Tumor-Segmentation-via-Neural-Shen/91455e43172bbc2fb1fa74adfd595d56fa6e7b68>
69. Poudel RPK, Lamata P, Montana G. Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation. In: Int Work Reconstr Anal Mov Body Organs 2016;83–94.
70. Cai J, Lu L, Xie Y, Xing F, Yang L. Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function. arXiv Prepr arXiv170704912 2017.
71. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–1780.
72. Jaccard P. The distribution of the flora in the alpine zone. New Phytol 1912;11:37–50.
73. Avendi MR, Kheradvar A, Jafarkhani H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. Med Image Anal 2016;30:108–119.
74. Rey D, Subsol G, Delingette H, Ayache N. Automatic detection and segmentation of evolving processes in 3D medical images: Application to multiple sclerosis. Med Image Anal 2002;6:163–179.
75. Roth HR, Lu L, Seff A, et al. A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Int Conf Med Image Comput Comput Interv 2014; 520–527.
76. Szegedy C, Reed S, Erhan D, Anguelov D, Ioffe S. Scalable, high-quality object detection. arXiv Prepr arXiv14121441 2014.
77. Erhan D, Szegedy C, Toshev A, Anguelov D. Scalable object detection using deep neural networks. In: Proc IEEE Conf Comput Vis Pattern Recognit 2014;2147–2154.
78. Dou Q, Chen H, Yu L, et al. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. IEEE Trans Med Imaging 2016;35:1182–1195.
79. Shin H-C, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35:1285–1298.
80. Samala RK, Chan H-P, Hadjiiski L, Helvie MA, Wei J, Cha K. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. Med Phys 2016;43:6654–6666.
81. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proc IEEE Conf Comput Vis pattern Recognit 2015;1–9.
82. Sa R, Owens W, Wiegand R, et al. Intervertebral disc detection in X-ray images using faster R-CNN. In: Eng Med Biol Soc (EMBC), 2017 39th Annu Int Conf IEEE 2017;564–567.
83. Liu J, Wang D, Lu L, et al. detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. Med Phys 2017;44:4630–4642.
84. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proc IEEE Conf Comput Vis Pattern Recognit 2016;779–788.
85. Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector. In: Eur Conf Comput Vis 2016;21–37.
86. Platania R, Shams S, Yang S, Zhang J, Lee K, Park S-J. Automated Breast Cancer Diagnosis Using Deep Learning and Region of Interest Detection (BC-DROID). In: Proc 8th ACM Int Conf Bioinformatics, Comput Biol Heal Informatics 2017;536–543.
87. Cao Z, Duan L, Yang G, et al. Breast tumor detection in ultrasound images using deep learning. In: Int Work Patch-based Tech Med Imaging 2017;121–128.
88. Li N, Liu H, Qiu B, et al. Detection and attention: Diagnosing pulmonary lung cancer from CT by imitating physicians. arXiv Prepr arXiv171205114 2017.
89. de Vos BD, Wolterink JM, de Jong PA, Viergever MA, Išgum I. 2D image classification for 3D anatomy localization: employing deep convolutional neural networks. In: Med Imaging Image Process 2016; 97841Y.
90. Prasoon A, Petersen K, Igel C, Lauze F, Dam E, Nielsen M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: Int Conf Med image Comput Comput Interv 2013; 246–253.



91. Roth HR, Lu L, Liu J, et al. Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Trans Med Imaging* 2016;35:1170–1181.
92. Roth HR, Wang Y, Yao J, Lu L, Burns JE, Summers RM. Deep convolutional networks for automated detection of posterior-element fractures on spine CT. *arXiv Prepr arXiv160200020* 2016.
93. Simonovsky M, Gutiérrez-Becker B, Mateus D, Navab N, Komodakis N. A deep metric for multimodal registration. In: *Int Conf Med Image Comput Comput Interv*. Berlin: Springer 2016;10–18.
94. Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P. Multimodality image registration by maximization of mutual information. *IEEE Trans Med Imaging* 1997;16.
95. Lv J, Yang M, Zhang J, Wang X. Respiratory motion correction for free-breathing 3D abdominal MRI using CNN based image registration: a feasibility study. *Br J Radiol* 2017;20170788.
96. de Vos BD, Berendsen FF, Viergever MA, Staring M, Išgum I. End-to-end unsupervised deformable image registration with a convolutional neural network BT — Deep learning in medical image analysis and multimodal learning for clinical decision support. Third International Workshop, DLMIA 2017, and 7th International. Edited by Cardoso MJ, Arbel T, Carneiro G, et al. Cham, Switzerland: Springer International Publishing 2017;204–212.
97. Bahrami K, Shi F, Zong X, Shin HW, An H, Shen D. Reconstruction of 7T-like images from 3T MRI. *IEEE Trans Med Imaging* 2016;35:2085–2097.
98. Schlemper J, Caballero J, Hajnal JV, Price A, Rueckert D. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Trans Med Imaging* 2017;PP:1.
99. Yang G, Yu S, Dong H, et al. DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE Trans Med Imaging* 2017;1–1.
100. Chatsias A, Joyce T, Giuffrida MV, Tsiftaris SA. Multimodal MR synthesis via modality-invariant latent representation. *IEEE Trans Med Imaging* 2017;62:1–1.
101. Chen H, Zhang Y, Zhang W, et al. Low-dose CT via convolutional neural network. *Biomed Opt Express* 2017;8:679–694.
102. Kang E, Min J, Ye JC. A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. *Med Phys* 2017;44:e360–e375.
103. Nie D, Cao X, Gao Y, Wang L, Shen D. Estimating CT image from MRI data using 3D fully convolutional networks BT — Deep learning and data labeling for medical applications. In: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICC. Edited by Carneiro G, Mateus D, Peter L, et al. Cham, Switzerland: Springer International Publishing 2016;170–178.
104. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med Phys* 2017;44:1408–1419.
105. Li R, Zhang W, Suk H-I, et al. Deep learning based imaging data completion for improved brain disease diagnosis. *Med Image Comput Assist Interv* 2014;17:305–312.
106. Oktay O, Bai W, Lee M, et al. Multi-input cardiac image super-resolution using convolutional neural networks. In: *Med Image Comput Comput Interv — MICCAI 2016 19th Int Conf Athens, Greece, Oct 17-21, 2016, Proceedings, Part III* 2016;246–254.
107. Umehara K, Ota J, Ishida T. Application of super-resolution convolutional neural network for enhancing image resolution in chest CT. *J Digit Imaging* 2017;1–10.
108. Chaudhari AS, Fang Z, Kogan F, et al. Super-resolution musculoskeletal MRI using deep learning. *Magn Reson Med* 2018 [Epub ahead of print].
109. Benou A, Veksler R, Friedman A, Riklin Raviv T. Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced MRI sequences. *Med Image Anal* 2017;42:145–159.
110. Jiang D, Dou W, Vosters L, Xu X, Sun Y, Tan T. Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network. *arxiv.org* 2017.
111. Gong E, Pauly JM, Wintermark M, Zaharchuk G. Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. *J Magn Reson Imaging* 2018 [Epub ahead of print].
112. Qayyum A, Anwar SM, Awais M, Majid M. Medical image retrieval using deep convolutional neural network. *Neurocomputing* 2017;266:8–20.
113. Chow LS, Paramesran R. Review of medical image quality assessment. *Biomed Signal Process Control* 2016;27:145–154.
114. Wu Lingyun, Cheng Jie-Zhi, Li Shengli, Lei Baiying, Wang Tianfu ND. FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks. *IEEE Trans Cybern* 2017;45:1336–1349.
115. Abdi AH, Luong C, Tsang T, et al. Automatic quality assessment of echocardiograms using convolutional neural networks: Feasibility on the apical four-chamber view. *IEEE Trans Med Imaging* 2017;36:1221–1230.
116. Esses SJ, Lu X, Zhao T, et al. Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture. *J Magn Reson Imaging* 2018;47:723–728.
117. Kim DH, Mackinnon T. *Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks*. Amsterdam: Elsevier 2017.
118. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *arXiv Prepr arXiv171005381* 2017.
119. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal* 2017;35:303–312.
120. Grewal M, Srivastava MM, Kumar P, Varadarajan S. RADNET: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans. *arXiv Prepr arXiv171004934* 2017.
121. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv Prepr arXiv171106504* 2017.
122. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology* 2017;170236.
123. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop* 2017;88:581–586.
124. Jamaludin A, Lootus M, Kadir T, et al. Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. *Eur Spine J* 2017;6:1374–1383.
125. Merkow J, Luftkin R, Nguyen K, Soatto S, Tu Z, Vedaldi A. DeepRadioNet: Radiologist level pathology detection in CT head images. *arXiv Prepr arXiv171109313* 2017.

## **A.2 A systematic study of the class imbalance problem in convolutional neural networks**



# A systematic study of the class imbalance problem in convolutional neural networks

Mateusz Buda<sup>a,b,\*</sup>, Atsuto Maki<sup>b</sup>, Maciej A. Mazurowski<sup>a,c</sup>

<sup>a</sup> Department of Radiology, Duke University School of Medicine, Durham, NC, USA

<sup>b</sup> School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

<sup>c</sup> Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA

## ARTICLE INFO

### Article history:

Received 17 January 2018

Received in revised form 26 May 2018

Accepted 20 July 2018

Available online 29 July 2018

### Keywords:

Class imbalance

Convolutional neural networks

Deep learning

Image classification

## ABSTRACT

In this study, we systematically investigate the impact of class imbalance on classification performance of convolutional neural networks (CNNs) and compare frequently used methods to address the issue. Class imbalance is a common problem that has been comprehensively studied in classical machine learning, yet very limited systematic research is available in the context of deep learning. In our study, we use three benchmark datasets of increasing complexity, MNIST, CIFAR-10 and ImageNet, to investigate the effects of imbalance on classification and perform an extensive comparison of several methods to address the issue: oversampling, undersampling, two-phase training, and thresholding that compensates for prior class probabilities. Our main evaluation metric is area under the receiver operating characteristic curve (ROC AUC) adjusted to multi-class tasks since overall accuracy metric is associated with notable difficulties in the context of imbalanced data. Based on results from our experiments we conclude that (i) the effect of class imbalance on classification performance is detrimental; (ii) the method of addressing class imbalance that emerged as dominant in almost all analyzed scenarios was oversampling; (iii) oversampling should be applied to the level that completely eliminates the imbalance, whereas the optimal undersampling ratio depends on the extent of imbalance; (iv) as opposed to some classical machine learning models, oversampling does not cause overfitting of CNNs; (v) thresholding should be applied to compensate for prior class probabilities when overall number of properly classified cases is of interest.

© 2018 Elsevier Ltd. All rights reserved.

## 1. Introduction

Convolutional neural networks (CNNs) are gaining significance in a number of machine learning application domains and are currently contributing to the state of the art in the field of computer vision, which includes tasks such as object detection, image classification, and segmentation. They are also widely used in natural language processing or speech recognition where they are replacing or improving classical machine learning models (Gu et al., 2015). CNNs integrate automatic feature extraction and discriminative classifier in one model, which is the main difference between them and traditional machine learning techniques. This property allows CNNs to learn hierarchical representations (Zeiler & Fergus, 2014). The standard CNN is built with fully connected layers and a number of blocks consisting of convolutions, activation function layer and max pooling (Krizhevsky, Sutskever, & Hinton, 2012; LeCun et al., 1989; Simonyan & Zisserman, 2014). The

complex nature of CNNs requires a significant computational power for training and evaluation of the networks, which is addressed with the help of modern graphical processing units (GPUs).

A common problem in real life applications of deep learning based classifiers is that some classes have a significantly higher number of examples in the training set than other classes. This difference is referred to as class imbalance. There are plenty of examples in domains like computer vision (Beijbom, Edmunds, Kline, Mitchell, & Kriegman, 2012; Johnson, Tateishi, & Hoan, 2013; Kubat, Holte, & Matwin, 1998; Van Horn et al., 2017; Xiao, Hays, Ehinger, Oliva, & Torralba, 2010), medical diagnosis (Grzymala-Busse, Goodwin, Grzymala-Busse, & Zheng, 2004; Mac Namee, Cunningham, Byrne, & Corrigan, 2002), fraud detection (Chan & Stolfo, 1998) and others (Cardie & Howe, 1997; Haixiang et al., 2016; Radivojac, Chawla, Dunker, & Obradovic, 2004) where this issue is highly significant and the frequency of one class (e.g., cancer) can be 1000 times less than another class (e.g., healthy patient). It has been established that class imbalance can have significant detrimental effect on training traditional classifiers (Japkowicz & Stephen, 2002) including multi-layer perceptrons (Mazurowski et al., 2008). It affects both convergence during the training phase and generalization of a model on the test set. While the issue very

\* Corresponding author at: Department of Radiology, Duke University School of Medicine, Durham, NC, USA

E-mail addresses: [buda@kth.se](mailto:buda@kth.se) (M. Buda), [atsuto@kth.se](mailto:atsuto@kth.se) (A. Maki), [maciej.mazurowski@duke.edu](mailto:maciej.mazurowski@duke.edu) (M.A. Mazurowski).

likely also affects deep learning, no systematic study on the topic is available.

Methods of dealing with imbalance are well studied for classical machine learning models (Chawla, 2005; Japkowicz & Stephen, 2002; Maloof, 2003; Mazurowski et al., 2008). The most straightforward and common approach is the use of sampling methods. Those methods operate on the data itself (rather than the model) to increase its balance. Widely used and proven to be robust is oversampling (Ling & Li, 1998). Another option is undersampling. Naïve version, called random majority undersampling, simply removes a random portion of examples from majority classes (Japkowicz & Stephen, 2002). The issue of class imbalance can be also tackled on the level of the classifier. In such case, the learning algorithms are modified, e.g. by introducing different weights to misclassification of examples from different classes (Zhou & Liu, 2006) or explicitly adjusting prior class probabilities (Lawrence, Burns, Back, Tsoi, & Giles, 1998).

Some previous studies showed results on cost sensitive learning of deep neural networks (Chung, Lin, & Yang, 2015; Khan, Benamoun, Sohel, & Togneri, 2015; Raj, Magg, & Wermter, 2016). New kinds of loss function for neural networks training were also developed (Wang et al., 2016). Recently, a new method for CNNs was introduced that trains the network in two-phases in which the network is trained on the balanced data first and then the output layers are fine-tuned (Havaei et al., 2017). While little systematic analysis of imbalance and methods to deal with it is available for deep learning, researchers employ some methods that might be addressing the problem likely based on intuition, some internal tests, and systematic results available for traditional machine learning. Based on our review of the literature, the method most commonly applied in deep learning is oversampling.

The remainder of this paper is organized as follows. Section 2 gives an overview of methods to address the problem of imbalance. In Section 3 we describe the experimental setup. It provides details about compared methods, datasets and models used for evaluation. Then, in Section 4 we present the results from our experiments and compare methods. Finally, Section 5 concludes the paper.

## 2. Methods for addressing imbalance

Methods for addressing class imbalance can be divided into two main categories (He & Garcia, 2009). The first category is data level methods that operate on training set and change its class distribution. They aim to alter dataset in order to make standard training algorithms work. The other category covers classifier (algorithmic) level methods. These methods keep the training dataset unchanged and adjust training or inference algorithms. Moreover, methods that combine the two categories are available. In this section we give an overview of commonly used approaches in both classical machine learning models and deep neural networks.

### 2.1. Data level methods

**Oversampling.** One of the most commonly used method in deep learning (Haixiang et al., 2016; Jaccard, Rogers, Morton, & Griffin, 2016; Janowczyk & Madabhushi, 2016; Levi & Hassner, 2015). The basic version of it is called random minority oversampling, which simply replicates randomly selected samples from minority classes. It has been shown that oversampling is effective, yet it can lead to overfitting (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Wang, Makond, Chen, & Wang, 2014). A more advanced sampling method that aims to overcome this issue is SMOTE (Chawla et al., 2002). It augments artificial examples created by interpolating neighboring data points. Some extensions of this technique were

proposed, for example focusing only on examples near the boundary between classes (Han, Wang, & Mao, 2005). Another type of oversampling approach uses data preprocessing to perform more informed oversampling. Cluster-based oversampling first clusters the dataset and then oversamples each cluster separately (Jo & Japkowicz, 2004). This way it reduces both between-class and within-class imbalance. DataBoost-IM, on the other hand, identifies difficult examples with boosting preprocessing and uses them to generate synthetic data (Guo & Viktor, 2004). An oversampling approach specific to neural networks optimized with stochastic gradient descent is class-aware sampling (Shen, Lin, & Huang, 2016). The main idea is to ensure uniform class distribution of each mini-batch and control the selection of examples from each class.

**Undersampling.** Another popular method (Haixiang et al., 2016) that results in having the same number of examples in each class. However, as opposed to oversampling, examples are removed randomly from majority classes until all classes have the same number of examples. While it might not appear intuitive, there is some evidence that in some situations undersampling can be preferable to oversampling (Drummond, Holte, et al., 2003). A significant disadvantage of this method is that it discards a portion of available data. To overcome this shortcoming, some modifications were introduced that more carefully select examples to be removed. E.g. one-sided selection identifies redundant examples close to the boundary between classes (Kubat, Matwin, et al., 1997). A more general approach than undersampling is data decontamination that can involve relabeling of some examples (Barandela, Rangel, Sánchez, & Ferri, 2003; Koplowitz & Brown, 1981).

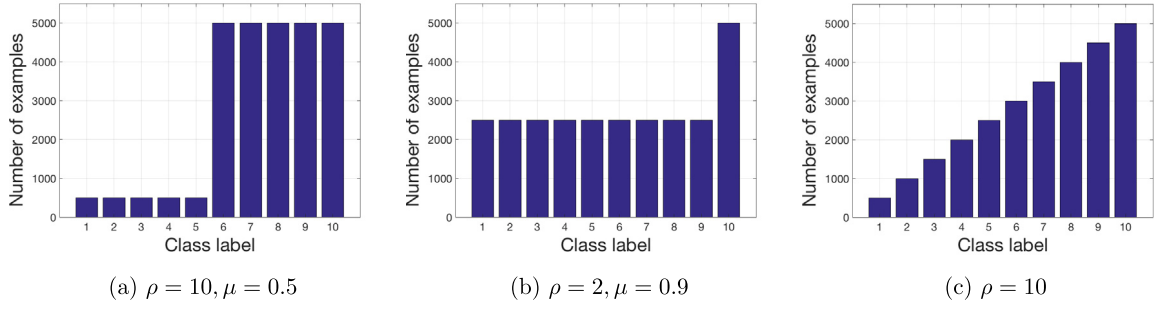
### 2.2. Classifier level methods

**Thresholding.** Also known as threshold moving or post scaling, adjusts the decision threshold of a classifier. It is applied in the test phase and involves changing the output class probabilities. There are many ways in which the network outputs can be adjusted. In general, the threshold can be set to minimize arbitrary criterion using an optimization algorithm (Lawrence et al., 1998). However, the most basic version simply compensates for prior class probabilities (Richard & Lippmann, 1991). These are estimated for each class by its frequency in the imbalanced dataset before sampling is applied. It was shown that neural networks estimate Bayesian a posteriori probabilities (Richard & Lippmann, 1991). That is, for a given datapoint  $x$ , their output for class  $i$  implicitly corresponds to

$$y_i(x) = p(i|x) = \frac{p(i) \cdot p(x|i)}{p(x)}.$$

Therefore, correct class probabilities can be obtained by dividing the network output for each class by its estimated prior probability  $p(i) = \frac{|i|}{\sum_k |k|}$ , where  $|i|$  denotes the number of unique examples in class  $i$ .

**Cost sensitive learning.** This method assigns different cost to misclassification of examples from different classes (Elkan, 2001). With respect to neural networks it can be implemented in various ways. One approach is threshold moving (Zhou & Liu, 2006) or post scaling (Lawrence et al., 1998) that is applied in the inference phase after the classifier is already trained. Similar strategy is to adapt the output of the network and also use it in the backward pass of backpropagation algorithm (Kukar, Kononenko, et al., 1998). Another adaptation of neural network to be cost sensitive is to modify the learning rate such that higher cost examples contribute more to the update of weights. And finally we can train the network by minimizing the misclassification cost instead of standard loss function (Kukar et al., 1998). The results of this approach are equivalent to oversampling (Chung et al., 2015; Zhou & Liu, 2006) described above and therefore this method will not be implemented in our study.



**Fig. 1.** Example distributions of imbalanced set together with corresponding values of parameters  $\rho$  and  $\mu$  for *step imbalance* (a–b) and  $\rho$  for *linear imbalance* (c).

**One-class classification.** In the context of neural networks it is usually called novelty detection. This is a concept learning technique that recognizes positive instances rather than discriminating between two classes. Autoencoders used for this purpose are trained to perform autoassociative mapping, i.e. identity function. Then, the classification of a new example is made based on a reconstruction error between the input and output patterns, e.g. absolute error, squared sum of errors, Euclidean or Mahalanobis distance (Japkowicz, Hanson, & Gluck, 2000; Japkowicz, Myers, Gluck, et al., 1995; Sohn, Worden, & Farrar, 2001). This method has proved to work well for extremely high imbalance when classification problem turns into anomaly detection (Lee & Cho, 2006).

**Hybrid of methods.** This is an approach that combines multiple techniques from one or both abovementioned categories. Widely used example is ensembling. It can be viewed as a wrapper to other methods. *EasyEnsemble* and *BalanceCascade* are methods that train a committee of classifiers on undersampled subsets (Liu, Wu, & Zhou, 2009). SMOTEBoost, on the other hand, is a combination of boosting and SMOTE oversampling (Chawla, Lazarevic, Hall, & Bowyer, 2003). Recently introduced and successfully applied to CNN training for brain tumor segmentation, is two-phase training (Havaei et al., 2017). Even though the task was image segmentation, it was approached as a pixel level classification. The method involves network pre-training on balanced dataset and then fine-tuning the last output layer before softmax on the original, imbalanced data.

### 3. Experiments

#### 3.1. Forms of imbalance

Class imbalance can take many forms particularly in the context of multiclass classification, which is typical in CNNs. In some problems only one class might be underrepresented or overrepresented and in other every class will have a different number of examples. In this study we define and investigate two types of imbalance that we believe are representatives of most of the real-world cases.

The first type is *step imbalance*. In *step imbalance*, the number of examples is equal within minority classes and equal within majority classes but differs between the majority and minority classes. This type of imbalance is characterized by two parameters. One is the fraction of minority classes defined by

$$\mu = \frac{|\{i \in \{1, \dots, N\} : C_i \text{ is minority}\}|}{N}, \quad (1)$$

where  $C_i$  is a set of examples in class  $i$  and  $N$  is the total number of classes. The other parameter is a ratio between the number of examples in majority classes and the number of examples in minority classes defined as follows.

$$\rho = \frac{\max_i \{|C_i|\}}{\min_i \{|C_i|\}} \quad (2)$$

An example of this type of imbalance is the situation when among the total of 10 classes, 5 of them have 500 training examples and another 5 have 5000. In this case  $\rho = 10$  and  $\mu = 0.5$ , as shown in Fig. 1a. A dataset with the same number of examples in total that has smaller imbalance ratio, corresponding to parameter  $\rho = 2$ , but more classes being minority,  $\mu = 0.9$ , is presented in Fig. 1b.

The second type of imbalance we call *linear imbalance*. We define it with one parameter that is a ratio between the maximum and minimum number of examples among all classes, as in Eq. (2) for imbalance ratio in *step imbalance*. However, the number of examples in the remaining classes is interpolated linearly such that the difference between consecutive pairs of classes is constant. An example of linear imbalance distribution with  $\rho = 10$  is shown in Fig. 1c.

#### 3.2. Methods of addressing imbalance compared in this study

In total, we examine seven methods to handle CNN training on a dataset with class imbalance which cover most of the commonly used approaches in the context of deep learning:

1. Random minority oversampling
2. Random majority undersampling
3. Two-phase training with pre-training on randomly oversampled dataset
4. Two-phase training with pre-training on randomly undersampled dataset
5. Thresholding with prior class probabilities
6. Oversampling with thresholding
7. Undersampling with thresholding

We examine two variants of two-phase training method. One on oversampled and the other on undersampled dataset. For the second phase, we keep the same hyperparameters and learning rate decay policy as in the first phase. Only the base learning rate from the first phase is multiplied by the factor of  $10^{-1}$ . Regarding thresholding, this method originally uses the imbalanced training set to train a neural network. We, in addition, combine it with oversampling and undersampling.

Selected methods are representative of the available approaches. Sampling can be used to explicitly incorporate cost of the examples by their appearance. It makes them one of many implementations of cost-sensitive learning (Zhou & Liu, 2006). Thresholding is another way of applying cost-sensitiveness by moving the output threshold such that higher cost examples are harder to misclassify. Ensemble methods require training of multiple classifiers. Because of considerable time needed to train deep models, it is often not practical and may be even infeasible to train multiple deep neural networks. One-class methods have a very limited application to datasets with extremely high imbalance. Moreover, they are applied to anomaly detection problem that is beyond the scope of our study.



**Table 1**

Summary of the used datasets. The number of images per class refers to the perfectly balanced subsets used for experiments. Provided image dimensions for ImageNet are given after rescaling.

Dataset	Image dimensions			No. classes	Images per class		CNN model
	Width	Height	Depth		Training	Test	
MNIST	28	28	1	10	5000	1000	LeNet-5
CIFAR-10	32	32	3	10	5000	1000	All-CNN
ILSVRC-2012	$\geq 256$	$\geq 256$	3	1000	1000	50	ResNet-10

**Table 2**

Architecture of LeNet-5 CNN used in MNIST experiments.

Layer	Data dimensions			Kernel size	Stride
	Width	Height	Depth		
Input	28	28	1	–	–
Convolution	24	24	20	5	1
Max pooling	12	12	20	2	2
Convolution	8	8	50	5	1
Max pooling	4	4	50	2	2
Fully connected	1	1	500	–	–
ReLU	1	1	500	–	–
Fully connected	1	1	10	–	–
Softmax	1	1	10	–	–

Importantly, we focused on methods that are widely used and relatively straightforward to implement as our aim is to draw conclusions that will be practical and serve as a guidance to a large number of deep learning researchers and engineers.

### 3.3. Datasets and models

In our study, we used three benchmark datasets: MNIST (LeCun, Bottou, Bengio, & Haffner, 1998), CIFAR-10 (Krizhevsky & Hinton, 2009) and ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 (Russakovsky et al., 2015). All of them are provided with a split on training and test set that are both labeled. For each dataset we choose different model with a set of hyperparameters used for its training that is known to perform well based on the literature. Datasets together with their corresponding models are of increasing complexity. This allows us to draw some conclusions on simple task and then verify how they scale to more complex ones.

All networks for the same dataset were trained with equal number of iterations. It means that the number of epochs differs between the imbalanced versions of dataset. This way we keep the number of weights' updates constant. Also, all networks were trained from a random initialization of weights and no pretraining was applied. An overview of some information about the datasets and their corresponding models is given in Table 1. All experiments were implemented in the deep learning framework Caffe (Jia et al., 2014).

#### 3.3.1. MNIST

MNIST is considered simple and solved problem that involves digits' images classification. The dataset consists of grayscale images of size  $28 \times 28$ . There are ten classes corresponding to digits from 0 to 9. The number of examples per class in the original training dataset ranges from 5421 in class 5 to 6742 in class 1. In artificially imbalanced versions we uniformly at random subsample each class to contain no more than 5000 examples.

The CNN model that we use for MNIST is the modern version of LeNet-5 (LeCun et al., 1998). The network architecture is presented in Table 2. All networks for this dataset were trained for 10 000 iterations. Optimization algorithm is stochastic gradient descent (SGD) with momentum value of  $\mu = 0.9$  (Qian, 1999). The learning rate decay policy is defined as  $\eta_t = \eta_0 \cdot (1 + \gamma \cdot t)^{-\alpha}$ , where  $\eta_0 = 0.01$  is a base learning rate,  $\gamma = 0.0001$  and  $\alpha = 0.75$  are decay

parameters and  $t$  is the current iteration. Furthermore, we used a batch size of 64 and a weight decay value of  $\lambda = 0.0005$ . Network weights were initialized randomly with uniform distribution and Xavier variance (Glorot & Bengio, 2010) whereas the biases were initialized with zero. No data augmentation was used. Test error of the model trained as described above on the original MNIST dataset was below 1%.

Experiments on MNIST dataset are performed on the following imbalance parameters space. For *linear imbalance* we test values of  $\rho \in \{10, 25, 50, 100, 250, 500, 1000, 2500, 5000\}$ . For *step imbalance* the set of  $\rho$  values is the same and for each we use all possible number of minority classes from 1 to 9, which corresponds to  $\mu \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ . The experiment for each combination of parameters is repeated 50 times. Every time the subset of minority classes is randomized. This way, we have created 4050 artificially imbalanced training sets for *step imbalance* and 450 for linear imbalance. As we have evaluated four methods that require training a model, the total number of trained networks, including baseline, is 22 500.

#### 3.3.2. CIFAR-10

CIFAR-10 is a significantly more complex image classification problem than MNIST. It contains  $32 \times 32$  color images with ten classes of natural objects. It does not have any natural imbalance at all. There are exactly 5000 training and 1000 test examples in each class. We do not use any data augmentation but follow standard preprocessing comprising global contrast normalization and ZCA whitening (Goodfellow, Warde-Farley, Mirza, Courville, & Bengio, 2013).

For CIFAR-10 experiments we use one of the best performing type of CNN model on this dataset, i.e. All-CNN (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014). The network architecture is presented in Table 3. The networks were trained for 70 000 iterations using SGD with momentum  $\mu = 0.9$ . The base learning rate was multiplied by a fixed multiplier of 0.1 after 40 000, 50 000 and 60 000 iterations. The number of examples in a batch was 256 and a weight decay value was  $\lambda = 0.001$ . Network weights were initialized with Xavier procedure and the biases set to zero. Test error of the model trained as described above on the original CIFAR-10 dataset was 9.75%.

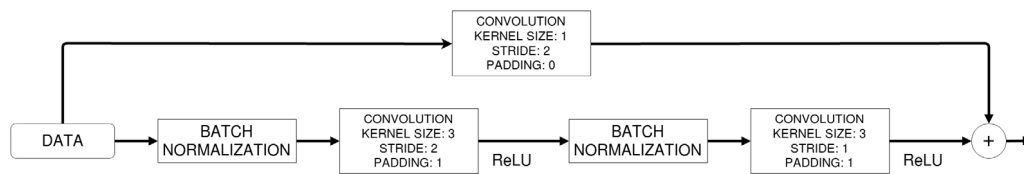
We have found the network training to be quite sensitive to initialization and the choice of base learning rate. Sometimes the network gets stuck in a very poor local minimum. Also, for more imbalanced datasets the training required lower base learning rate to train at all. Therefore, for each case we were searching for the best one from the fixed set  $\eta_0 \in \{0.05, 0.005, 0.0005, 0.00005\}$ . Similar procedure was used by the authors of the model architecture (Springenberg et al., 2014). Moreover, each training was repeated twice on the same dataset. For a particular method and imbalanced dataset, we pick the model with the best score on the test set over all eight runs.

The network architecture does not have any fully connected layers. Therefore, during the fine-tuning in two-phase training method we update the weights of two last convolutional layers with kernels of size 1.

The imbalance parameters space used in CIFAR-10 experiments is considerably sparser than the one used for MNIST due to the

**Table 3**  
Architecture of All-CNN used in CIFAR-10 experiments.

Layer	Data dimensions			Kernel size	Stride	Padding
	Width	Height	Depth			
Input	32	32	3	–	–	–
Dropout (0.2)	32	32	3	–	–	–
2×(Convolution + ReLU)	32	32	96	3	1	1
Convolution + ReLU	16	16	96	3	2	1
Dropout (0.5)	16	16	96	–	–	–
2×(Convolution + ReLU)	16	16	192	3	1	1
Convolution + ReLU	8	8	192	3	2	1
Dropout (0.5)	8	8	192	–	–	–
Convolution + ReLU	6	6	192	3	1	0
Convolution + ReLU	6	6	192	1	1	0
Convolution + ReLU	6	6	10	1	1	0
Average Pooling	1	1	10	6	–	–
Softmax	1	1	10	–	–	–



**Fig. 2.** Architecture of a single residual block in ResNet used in ILSVRC-2012 experiments.

significantly longer time required to train one network. The set of tested values was narrowed to make the experiment run in a reasonable time. For *linear* and *step imbalance*, we test values of  $\rho \in \{2, 10, 20, 50\}$ . In *step imbalance*, for each value of  $\rho$ , the set of values of parameter  $\mu$  was  $\mu \in \{0.2, 0.5, 0.8\}$ , which corresponds to having two, five and eight minority classes, respectively. And for all the cases, the classes chosen to be minority were the ones with the lowest label value. It means that for a fixed number of minority classes the same classes were always picked as minority. Also, all of them were included in a larger set of minority classes. In total we trained 640 networks on this dataset.

### 3.3.3. ImageNet

For evaluation we use a ILSVRC-2012 competition subset of ImageNet, widely used as a benchmark to compare classifiers' performance. The number of examples in majority classes was reduced from 1200 to 1000. Classes with less than 1000 cases were always chosen as a minority ones for imbalanced subsets. The only data preprocessing applied is resizing such that the smaller dimension is 256 pixels long and the aspect ratio is preserved. During training, as input we use a randomly cropped  $224 \times 224$  pixel square patch and a single centered crop in a test phase. Moreover, during training we randomly mirror images, but there is no color, scale or aspect ratio augmentation.

A model architecture employed for this dataset is ResNet-10 (Simon, Rodner, & Denzler, 2016), i.e. a residual network (He, Zhang, Ren, & Sun, 2016) with batch normalization layers that are known to accelerate deep networks training (Ioffe & Szegedy, 2015). It consists of four residual blocks that give us nine convolutional layers and one fully connected. The first residual block outputs data tensor of depth 64 and then each one increases it by a factor of two. Fully connected layer outputs 1000 values to softmax that transforms them to class probabilities. The architecture of one residual block is presented in Fig. 2.

The networks were trained for 320 000 iterations using SGD with momentum  $\mu = 0.9$ . The base learning rate is set to  $\eta_0 = 0.1$  and decays linearly to 0 in the last iteration. The number of examples in a batch was 256 and a weight decay  $\lambda = 0.0001$ . Network weights were initialized with Kaiming (also known as MSRA) initialization procedure (He, Zhang, Ren, & Sun, 2015). Top-1 test error of the model trained as described above on the original

ILSVRC-2012 dataset was 62.56% and 99.50 multi-class ROC AUC for a single centered crop.

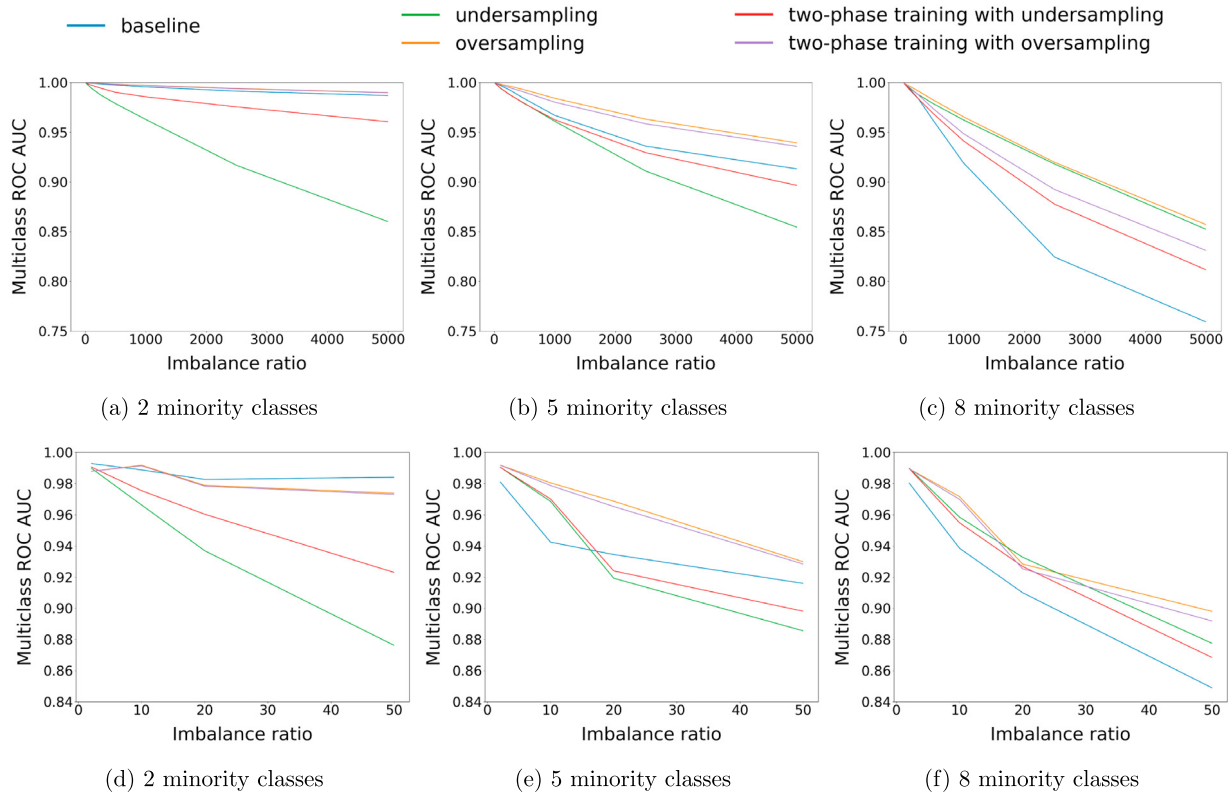
We have chosen relatively small ResNet for the sake of faster training but without loss of generality.<sup>1</sup> We test only one case of small and two cases of large *step imbalance* and run it on the baseline, undersampling and oversampling methods. Specifically, all three *step imbalanced* subsets are defined with  $\mu = 0.1$ ,  $\rho = 10$ ,  $\mu = 0.8$ ,  $\rho = 50$  and  $\mu = 0.9$ ,  $\rho = 100$ . They correspond to 100 minority classes with imbalance ratio of 10, 800 minority classes with imbalance of 50, and 900 minority classes with imbalance ratio of 100, respectively. Moreover, for the highest imbalance, we train three networks for each method with randomized selection of minority classes and subsampled set of examples in each class. This is done in order to estimate variability in performance of methods. In total, this gives us 15 ResNet-10 networks trained on five artificially imbalanced subsets of ILSVRC-2012.

### 3.4. Evaluation metrics and testing

The metric that is most widely used to evaluate a classifier performance in the context of multiclass classification with CNNs is overall accuracy which is the proportion of test examples that were correctly classified. However, it has some significant and long acknowledged limitations, particularly in the context of imbalanced datasets (Chawla, 2005). Specifically, when the test set is imbalanced, accuracy will favor classes that are overrepresented in some cases leading to highly misleading assessment. An example of this is a situation when the majority class represents 99% of all cases and the classifier assigns the label of the majority class to all test cases. A misleading accuracy of 99% will be assigned to a classifier that has a very limited use. Another issue might arise when the test set is balanced and a training set is imbalanced. This might result in a situation when a decision threshold is moved to reflect the estimated class prior probabilities and cause a low accuracy measure in the test set while the true discriminative power of the classifier does not change.

A measure that addresses these issues is area under the receiver operating characteristic curve (ROC AUC) (Bradley, 1997) which is a

<sup>1</sup> It takes five days to train one ResNet-10 network on Nvidia GTX 1070 GPU.



**Fig. 3.** Comparison of methods with respect to multi-class ROC AUC on MNIST (a–c) and CIFAR-10 (d–f) for step imbalance with fixed number of minority classes.

plot of the false positive rate to the true positive rate for all possible prediction thresholds. We used a specific implementation of the ROC AUC available in scikit-learn python package (Pedregosa et al., 2011). It calculates sensitivities and specificities at all thresholds defined by the responses of the classifier in the test set followed by the AUC calculation using the trapezoid rule. ROC AUC is a well-studied and sound measure of discrimination (Ling, Huang, & Zhang, 2003) and has been widely used as an evaluation metric for classifiers. ROC has also been used to compare performance of classifiers trained on imbalanced datasets (Maloof, 2003; Mazurowski et al., 2008). Since the basic version of ROC is only suitable for binary classification, we use a multi-class modification of it (Provost & Domingos, 2003). The multi-class ROC is calculated by taking the average of AUCs obtained independently for each class for the binary classification task of distinguishing a given class from all the other classes.

Test set of all used datasets has equal number of examples in each class. Usually, it is assumed that the class distribution of a test set follows the one of a training set. We do not change a test set to match artificially imbalanced training set. The reason is that the score achieved by each classifier on the same test set is more comparable and the largest number of cases in each of the classes provides the most accurate performance estimation.

## 4. Results

### 4.1. Effects of class imbalance on classification performance and comparison of methods to address imbalance

The results showing the impact of class imbalance on classification performance and comparison of methods for addressing imbalance are shown in Figs. 3 and 4. Fig. 3 shows the results with respect to multi-class ROC AUC for a fixed number of minority classes on MNIST and CIFAR-10. Fig. 4 presents the result from the

perspective of fixed ratio of imbalance, i.e. parameter  $\rho$ , for the same two datasets.

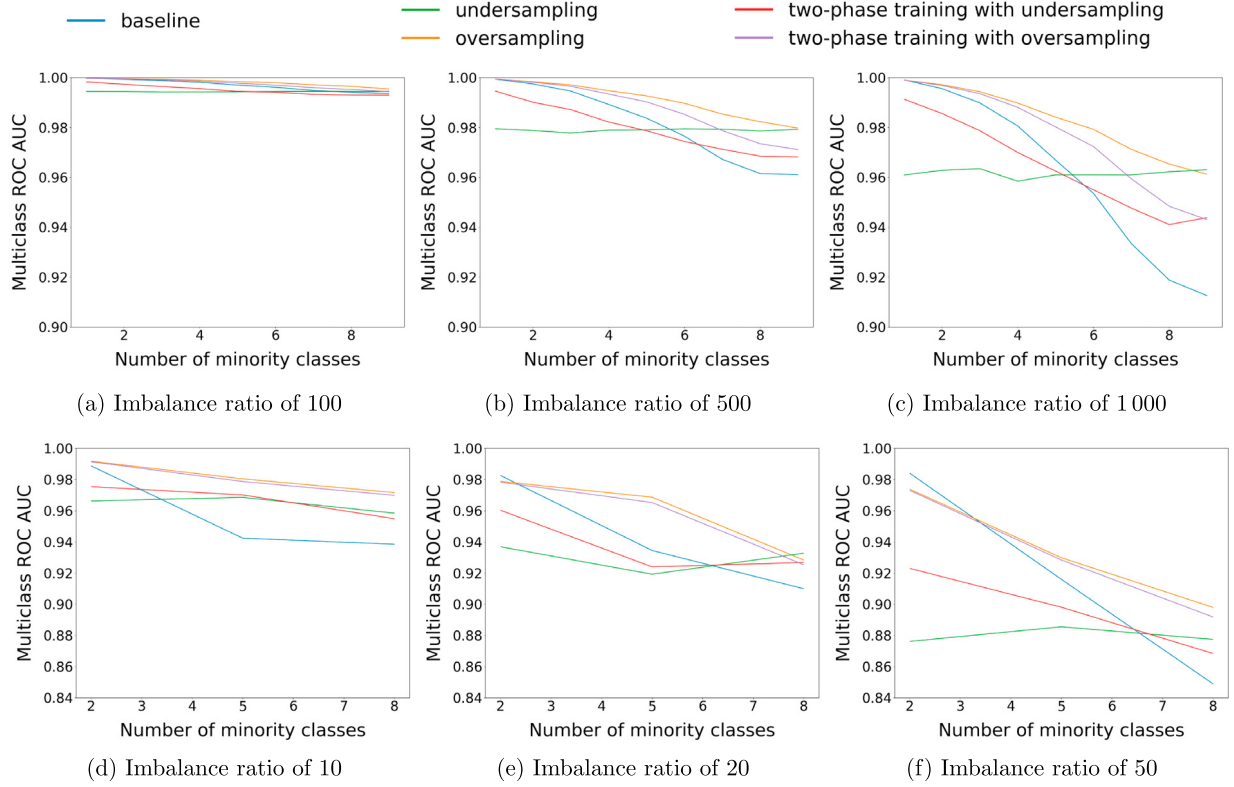
Regarding the effect of class imbalance on classification performance, we observed the following. First, the deterioration of performance due to class imbalance is substantial. As expected, the increasing ratio of examples between majority and minority classes as well as the number of minority classes had a negative effect on performance of the resulting classifiers. Furthermore, by comparing the results from MNIST and CIFAR-10 we observed that the effect of imbalance is significantly stronger for the task with higher complexity. A similar drop in performance for MNIST and CIFAR-10 corresponded to approximately 100 times stronger level of imbalance in the MNIST dataset.

Regarding performance of different methods for addressing imbalance, in almost all of the situations oversampling emerged as the best method. It also showed notable improvement of performance over the baseline (i.e. do-nothing strategy) in majority of the situations and never showed a considerable decrease in performance for the two datasets analyzed in this section making it a clear recommendation for tasks similar to MNIST and CIFAR-10.

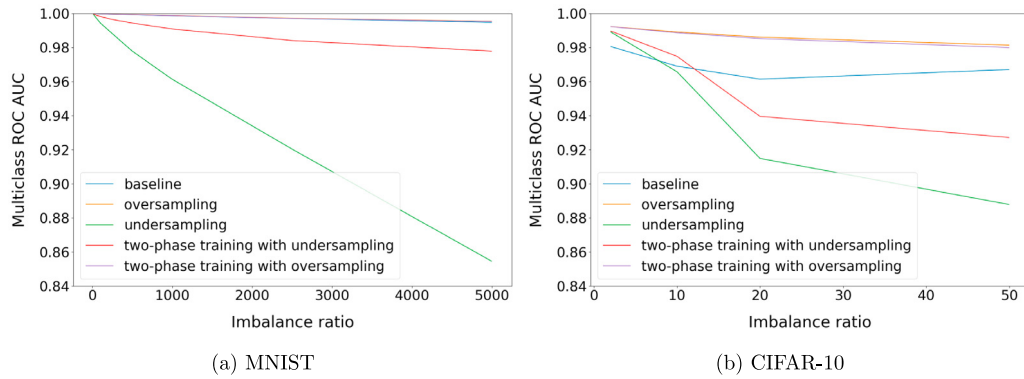
Undersampling showed a generally poor performance. In a large number of analyzed scenarios undersampling showed decrease in performance as compared to the baseline. In scenarios with a large proportion of minority classes undersampling showed some improvement over the baseline but never a notable advantage over oversampling (Fig. 3).

For a fixed imbalance ratio undersampling is always trained on the subset of equal size. As a result, its performance does not change with the number of minority classes. For both datasets and each case of imbalance ratio, the gap between undersampling and oversampling is the biggest for smaller number of minority classes and decreases with the number of minority classes, as shown in Fig. 4. This is expected since with all classes being minority these two methods become equivalent.





**Fig. 4.** Comparison of methods with respect to multi-class ROC AUC on MNIST (a–c) and CIFAR-10 (d–f) for *step imbalance* with fixed imbalance ratio.



**Fig. 5.** Comparison of methods with respect to multi-class ROC AUC for *linear imbalance*.

Two-phase training methods with both undersampling and oversampling tend to perform between the baseline and their corresponding method (undersampling or oversampling). If the baseline is better than one of these methods, fine-tuning improves the original method. Otherwise, performance deteriorates. However, if the baseline is better, there is still no gain from using two-phase training method. As oversampling is almost always better than the baseline, fine-tuning always gives lower score.

The variability of patterns, visual structures, and objects in CIFAR-10 is considerably higher than in MNIST. For this reason, we run the *step imbalance* experiment three times on a reshuffled stratified training and test split to validate our results. Additional results are available in Appendix A.

In Fig. 5 we show the results for *linear imbalance* on MNIST and CIFAR-10 datasets. The highest possible *linear imbalance* ratio for MNIST dataset is 5 000, which means only one example in the most underrepresented class. However, even in this case the decrease in performance according to multi-class ROC AUC score for the

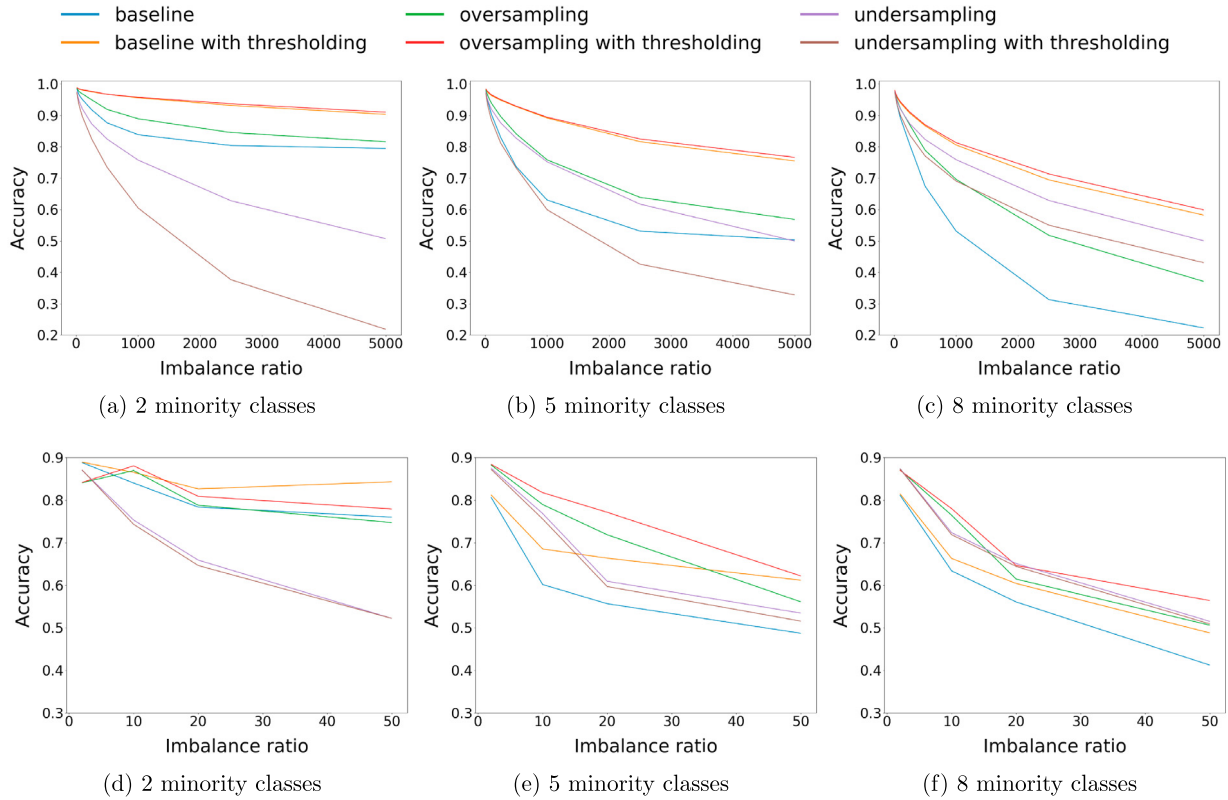
baseline model is not significant, as shown in Fig. 5a. Nevertheless, oversampling improves the score on both datasets and for all tested values of  $\rho$ , whereas the score for undersampling decreases approximately linearly with imbalance ratio.

#### 4.2. Results on ImageNet dataset

The results from experiments performed on ImageNet (ILSVRC-2012) dataset confirm the impact of imbalance on classifier's performance. Table 4 compares methods with respect to multi-class ROC AUC. The drop in performance for the largest tested imbalance was from 99 to 90, in terms of multi-class ROC AUC. The results confirm that the oversampling approach performs consistently better than undersampling approach across all scenarios. A small decrease in performance as compared to baseline was observed for oversampling for extreme imbalances. Please note, however, that these results should be treated with caution and not as strong

**Table 4**  
Comparison of results on ImageNet with respect to multi-class ROC AUC.

Method	$\mu = 0.1, \rho = 10$	$\mu = 0.8, \rho = 50$	$\mu = 0.9, \rho = 100$		
Baseline	99.41	96.31	90.74	90.46	90.05
Oversampling	99.35	95.06	88.38	88.39	88.17
Undersampling	96.85	94.98	88.35	84.08	83.74



**Fig. 6.** Comparison of methods with respect to accuracy on MNIST (a–c) and CIFAR-10 (d–f) for *step imbalance* with fixed number of minority classes.

evidence that oversampling is inferior for highly complex tasks with extreme imbalance. The absolute difference in performance between three runs with respect to multi-class ROC AUC was even higher than 4 (for undersampling). Therefore, differences of 1–2 might be due to variability of results between different runs of neural networks. Moreover, the highest tested imbalanced training set was only about 10% of the original ILSVRC-2012 introducing confounding issues such as the optimal training hyperparameters for this significantly changed dataset. Therefore, while these results indicate that caution should be taken when any sampling technique is applied to highly complex tasks with extreme imbalances, it needs a more extensive study devoted to this specific issue.

#### 4.3. Separation of effects from reduced number of examples and class imbalance

An important question that needs to be considered in the context of our study is whether the decrease in performance for imbalanced datasets is merely caused by the fact that our imbalanced datasets simply had fewer training examples or is it truly caused by the fact that the datasets are imbalanced.

First, we notice that oversampling method uses the same amount of data as the baseline. It only eliminates the imbalance which is enough to improve the performance in almost all the cases. Still, it does not reach the performance of a classifier trained

on the original dataset. This is an indication that the effect of imbalance is not trivial.

Second, for some cases undersampling, which reduces the total number of cases performs better than the baseline (see Figs. 3c and 3f). Moreover, there are even cases when undersampling can perform on a par with oversampling. It means that, between two sampling methods that eliminate imbalance, even using fewer data can be comparable.

In addition, for the same value of parameter  $\rho$  we have equal number of examples in the training set for *linear imbalance* and *step imbalance* with  $\mu = 0.5$ , which corresponds to half of the classes being minority. The drop in performance is much higher for *step imbalance*. This additionally demonstrates that not only the total number of examples matters but also its distribution between classes.

#### 4.4. Improving accuracy score with multi-class thresholding

While our focus is on ROC AUC, we also provide the evaluation of the methods based on overall accuracy measure with results on *step imbalance* shown in Fig. 6. As explained in Section 3.4, accuracy has some known limitations and in some scenarios does not reflect the discriminative power of a classifier but rather the prevalence of classes in the training or test set. Nevertheless, it is still commonly used evaluation score (Haixiang et al., 2016) and therefore we provide some results according to this metric.

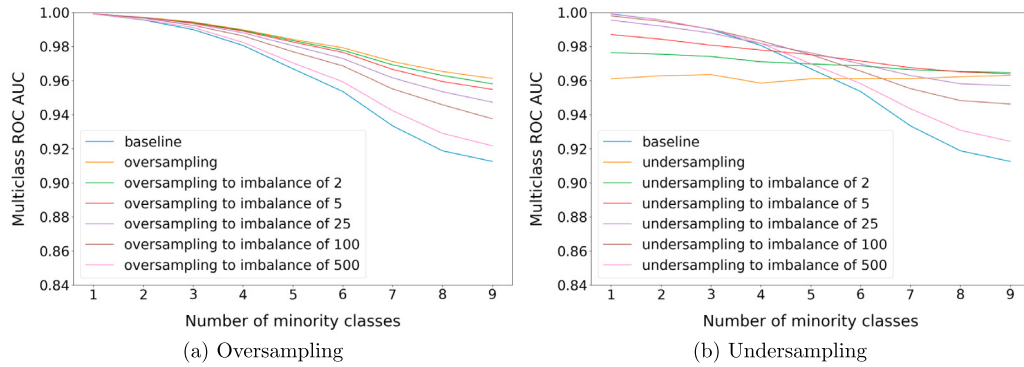


Fig. 7. Comparison of oversampling and undersampling to reduced imbalance ratios on MNIST with original imbalance of 1000.

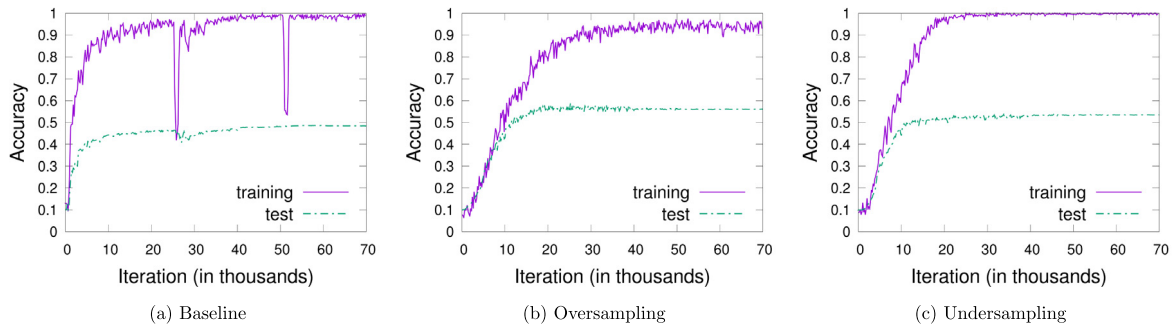


Fig. 8. Comparison of networks convergence between baseline, oversampling and undersampling. Training on CIFAR-10 *step imbalanced* with 5 minority classes and imbalance ratio of 50.

Our results show that thresholding is an appropriate approach to take to offset the prior probabilities of different classes learned by a network based on imbalanced datasets and provided an improvement in overall accuracy. In general, thresholding worked particularly well when applied jointly with oversampling.

Please note that thresholding does not have an actual effect on the ability of the classifier to discriminate between a given class from another but rather helps to find a threshold on the network output that guarantees a large number of correctly classified cases. In terms of ROC, multiplying a decision variable by any positive number does not change the area under the ROC curve. However, finding an optimal operating point on the ROC curve is important when the overall number of correctly classified cases is of interest.

#### 4.5. Undersampling and oversampling to smaller imbalance ratio

The default version of oversampling is to increase the number of cases in the minority classes so that the number matches the majority classes. Similarly, the default of undersampling is to decrease the number of cases in the majority classes to match the minority classes. However, a more moderate version of these algorithms could be applied. For the case of MNIST with imbalance ratio of 1000 we have tried to gradually decrease the imbalance with oversampling and undersampling. The results are shown in Fig. 7.

The results show that the default version of oversampling was always the best. Any reduction of imbalance improves the score regardless of the number of minority classes, as shown in Fig. 7a. For undersampling, in some cases of moderate number of minority classes, intermediate levels of undersampling performed better than both full undersampling and the baseline.

Moreover, comparing undersampling and oversampling to reduced level of imbalance, we can notice that for each case of oversampling there is a level to which we can apply undersampling

and achieve equivalent performance. However, that level is not known a priori rendering oversampling still the method of choice.

#### 4.6. Generalization of sampling methods

In some cases undersampling and oversampling perform similarly. In those cases, one would probably prefer the model that generalizes better. For classical machine learning models it was shown that oversampling can cause overfitting, especially for minority classes (Chawla et al., 2002). As we repeat small number of examples multiple times, the trained model fits them too well. Thus, according to this prior knowledge undersampling would be a better choice. The results from our experiments do not confirm this conclusion for convolutional neural networks.

In Fig. 8 we compare the convergence of baseline and sampling methods for CIFAR-10 experiments with respect to accuracy. Both oversampling and undersampling methods helped to train a better classifier in terms of performance and generalization. They also made training more stable. As opposed to traditional machine learning methods, in this case oversampling did not lead to overfitting. The gap between accuracy on the training and test set does not increase with iterations for oversampling, Fig. 8b. Furthermore, we validated this phenomenon in multiple additional scenarios for all analyzed datasets and have not observed overfitting in any of these scenarios. The additional plots are included in Appendix B.

### 5. Conclusions

In this study, we examined the impact of class imbalance on classification performance of convolutional neural networks and investigated the effectiveness of different methods of addressing the issue. We defined and parametrized two representative types of imbalance, i.e. step and linear. Then we subsampled MNIST,

CIFAR-10 and ImageNet (ILSVRC-2012) datasets to make them artificially imbalanced. We have compared common sampling methods, basic thresholding, and two-phase training.

The observations from our experiments related to the class imbalance are as follows.

- The effect of class imbalance on classification performance is detrimental.
- The influence of imbalance on classification performance increases with the scale of a task.
- The impact of imbalance cannot be explained simply by the lower total number of training cases and depends on the distribution of examples among classes.

Regarding the choice of a method to handle CNN training on imbalanced dataset we conclude the following.

- The method that in most of the cases outperforms all others with respect to multi-class ROC AUC was oversampling.
- For extreme ratio of imbalance and large portion of classes being minority, undersampling performs on a par with oversampling. If training time is an issue, undersampling is a better choice in such a scenario since it dramatically reduces the size of the training set.
- To achieve the best accuracy, one should apply thresholding to compensate for prior class probabilities. A combination of thresholding with baseline and oversampling is the most preferable, whereas it should not be combined with undersampling.
- Oversampling should be applied to the level that completely eliminates the imbalance, whereas the optimal undersampling ratio depends on the extent of imbalance. The higher a fraction of minority classes in the imbalanced training set, the more imbalance ratio should be reduced.
- Oversampling does not cause overfitting of convolutional neural networks, as opposed to some classical machine learning models.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.neunet.2018.07.011>.

## References

- Barandela, R., Rangel, E., Sánchez, J. S., & Ferri, F. J. (2003). Restricted decontamination for the imbalanced training sample problem. In *Iberoamerican congress on pattern recognition* (pp. 424–431). Springer.
- Beijbom, O., Edmunds, P. J., Kline, D. I., Mitchell, B. G., & Kriegman, D. (2012). Automated annotation of coral reef survey images. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 1170–1177). IEEE.
- Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145–1159.
- Cardie, C., & Howe, N. (1997). Improving minority class prediction using case-specific feature weights. In *ICML* (pp. 57–65).
- Chan, P. K., & Stolfo, S. J. (1998). Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *KDD, Vol. 1998* (pp. 164–168).
- Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook* (pp. 853–867). Springer.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery* (pp. 107–119). Springer.
- Chung, Y. A., Lin, H. T., & Yang, S. W. Cost-aware pre-training for multiclass cost-sensitive deep learning, arXiv preprint [arXiv:1511.09337](https://arxiv.org/abs/1511.09337).
- Drummond, C., & Holte, R. C. et al., (2009). C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II, Vol. 11* (pp. 1–8).
- Elkan, C. (2001). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence, Vol. 17* (pp. 973–978). Lawrence Erlbaum Associates Ltd.
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feed-forward neural networks. In *Aistats, Vol. 9* (pp. 249–256).
- Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A. C., & Bengio, Y. (2013). Maxout networks. In *ICML (3), Vol. 28* (pp. 1319–1327).
- Grzymala-Busse, J. W., Goodwin, L. K., Grzymala-Busse, W. J., & Zheng, X. (2004). An approach to imbalanced data sets based on changing rule strength. In *Rough neural computing* (pp. 543–553). Springer.
- Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., & Shuai, B. et al., Recent advances in convolutional neural networks, arXiv preprint [arXiv:1512.07108](https://arxiv.org/abs/1512.07108).
- Guo, H., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *ACM Sigkdd Explorations Newsletter*, 6(1), 30–39.
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2016). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*.
- Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. *Advances in Intelligent Computing*, 878–887.
- Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35, 18–31.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *The IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Ioffe, S., & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint [arXiv:1502.03167](https://arxiv.org/abs/1502.03167).
- Jaccard, N., Rogers, T. W., Morton, E. J., & Griffin, L. D. (2016). Detection of concealed cars in complex cargo X-ray imagery using deep learning. *Journal of X-Ray Science and Technology*, 1–17.
- Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7.
- Japkowicz, N., Hanson, S. J., & Gluck, M. A. (2000). Nonlinear autoassociation is not equivalent to pca. *Neural Computation*, 12(3), 531–545.
- Japkowicz, N., Myers, C., & Gluck, M. et al., (1995). A novelty detection approach to classification. In *IJCAI, Vol. 1* (pp. 518–523).
- Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429–449.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on multimedia* (pp. 675–678). ACM.
- Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter*, 6(1), 40–49.
- Johnson, B. A., Tateishi, R., & Hoan, N. T. (2013). A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. *International Journal of Remote Sensing*, 34(20), 6969–6982.
- Khan, S. H., Bennamoun, M., Soheli, F., & Togneri, R. Cost sensitive learning of deep feature representations from imbalanced data, arXiv preprint [arXiv:1508.03422](https://arxiv.org/abs/1508.03422).
- Koplowitz, J., & Brown, T. A. (1981). On the relation of performance to editing in nearest neighbor rules. *Pattern Recognition*, 13(3), 251–255.
- Krizhevsky, A., & Hinton, G. Learning multiple layers of features from tiny images. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).
- Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2–3), 195–215.
- Kubat, M., & Matwin, S. et al., (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *ICML, Vol. 97* (pp. 179–186). Nashville, USA.
- Kukar, M., & Kononenko, I. et al., (1998). Cost-sensitive learning with neural networks. In *ECAI* (pp. 445–449).
- Lawrence, S., Burns, I., Back, A., Tsoi, A. C., & Giles, C. L. (1998). Neural network classification and prior class probabilities. In *Neural networks: Tricks of the trade* (pp. 299–313). Springer.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4), 541–551.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.
- Lee, H. J., & Cho, S. (2006). The novelty detection approach for different degrees of class imbalance. In *Neural information processing* (pp. 21–30). Springer.



- Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 34–42).
- Ling, C. X., Huang, J., & Zhang, H. (2003). Auc: a statistically consistent and more discriminating measure than accuracy. In *IJCAI, Vol. 3* (pp. 519–524).
- Ling, C. X., & Li, C. (1998). Data mining for direct marketing: Problems and solutions. In *KDD, Vol. 98* (pp. 73–79).
- Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 39(2), 539–550.
- Mac Namee, B., Cunningham, P., Byrne, S., & Corrigan, O. I. (2002). The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, 24(1), 51–70.
- Maloof, M. A. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML-2003 workshop on learning from imbalanced data sets II*.
- Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2), 427–436.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830.
- Provost, F., & Domingos, P. (2003). Tree induction for probability-based ranking. *Machine Learning*, 52(3), 199–215.
- Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1), 145–151.
- Radivojac, P., Chawla, N. V., Dunker, A. K., & Obradovic, Z. (2004). Classification and knowledge discovery in protein databases. *Journal of Biomedical Informatics*, 37(4), 224–239.
- Raj, V., Magg, S., & Wermter, S. (2016). Towards effective classification of imbalanced data with convolutional neural networks. In *IAPR workshop on artificial neural networks in pattern recognition* (pp. 150–162). Springer.
- Richard, M. D., & Lippmann, R. P. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, 3(4), 461–483.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.
- Shen, L., Lin, Z., & Huang, Q. (2016). Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision* (pp. 467–482). Springer.
- Simon, M., Rodner, E., & Denzler, J. Imagenet pre-trained models with batch normalization, arXiv preprint [arXiv:1612.01452](https://arxiv.org/abs/1612.01452).
- Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition, arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- Sohn, H., Worden, K., & Farrar, C. R. (2001). Novelty detection using auto-associative neural network. In *Symposium on identification of mechanical systems: international mechanical engineering congress and exposition* (pp. 573–580), New York, NY.
- Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. Striving for simplicity: The all convolutional net, arXiv preprint [arXiv:1412.6806](https://arxiv.org/abs/1412.6806).
- Van Horn, G., Mac Aodha, O., Song, Y., Shepard, A., Adam, H., & Perona, P. et al., The inaturalist challenge 2017 dataset, arXiv preprint [arXiv:1707.06642](https://arxiv.org/abs/1707.06642).
- Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., & Kennedy, P. J. (2016). Training deep neural networks on imbalanced data sets. In *2016 international joint conference on neural networks* (pp. 4368–4374). IEEE.
- Wang, K. J., Makond, B., Chen, K. H., & Wang, K. M. (2014). A hybrid classifier combining smote with pso to estimate 5-year survivability of breast cancer patients. *Applied Soft Computing*, 20, 15–24.
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE conference on computer vision and pattern recognition* (pp. 3485–3492). IEEE.
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.
- Zhou, Z. H., & Liu, X. Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1), 63–77.

### **A.3 Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm**



# Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm

Mateusz Buda<sup>a,\*</sup>, Ashirbani Saha<sup>a</sup>, Maciej A. Mazurowski<sup>a,b</sup>

<sup>a</sup> Department of Radiology, Duke University School of Medicine, 2424 Erwin Road, Suite 302, Durham, NC, 27705, USA

<sup>b</sup> Department of Electrical and Computer Engineering, Duke University, Box 90291, Durham, NC, 27708, USA

## ARTICLE INFO

### Keywords:

Deep learning  
Brain segmentation  
Radiogenomics  
MRI  
LGG

## ABSTRACT

Recent analysis identified distinct genomic subtypes of lower-grade glioma tumors which are associated with shape features. In this study, we propose a fully automatic way to quantify tumor imaging characteristics using deep learning-based segmentation and test whether these characteristics are predictive of tumor genomic subtypes.

We used preoperative imaging and genomic data of 110 patients from 5 institutions with lower-grade gliomas from The Cancer Genome Atlas. Based on automatic deep learning segmentations, we extracted three features which quantify two-dimensional and three-dimensional characteristics of the tumors. Genomic data for the analyzed cohort of patients consisted of previously identified genomic clusters based on IDH mutation and 1p/19q co-deletion, DNA methylation, gene expression, DNA copy number, and microRNA expression. To analyze the relationship between the imaging features and genomic clusters, we conducted the Fisher exact test for 10 hypotheses for each pair of imaging feature and genomic subtype. To account for multiple hypothesis testing, we applied a Bonferroni correction. *P*-values lower than 0.005 were considered statistically significant.

We found the strongest association between RNASeq clusters and the bounding ellipsoid volume ratio ( $p < 0.0002$ ) and between RNASeq clusters and margin fluctuation ( $p < 0.005$ ). In addition, we identified associations between bounding ellipsoid volume ratio and all tested molecular subtypes ( $p < 0.02$ ) as well as between angular standard deviation and RNASeq cluster ( $p < 0.02$ ). In terms of automatic tumor segmentation that was used to generate the quantitative image characteristics, our deep learning algorithm achieved a mean Dice coefficient of 82% which is comparable to human performance.

## 1. Introduction

Lower-grade gliomas (LGG) are a group of WHO grade II and grade III brain tumors. As opposed to grade I which are often curable by surgical resection, grade II and III are infiltrative and tend to recur and evolve to higher-grade lesion. Predicting patient outcomes based on histopathological data for these tumors is inaccurate and suffers from inter-observer variability [1]. One of the promising methods that might address this issue is defining subtypes of LGG through clustering of patients based on DNA methylation, gene expression, DNA copy number, and microRNA expression [1]. It was shown that the clusters identified in such way are to a large extent in agreement with another basic molecular subtype based on IDH (IDH1 and IDH2) mutation and 1p/19q co-deletion [1,2]. Patients with tumors from different molecular groups substantially differ in terms of typical course of the disease and overall survival [3].

A new research direction in cancer, called radiogenomics, aims at investigating the relationship between tumor genomic characteristics and medical imaging [4]. Imaging can provide important information before surgery or in cases when resection is not possible. Very recent studies in this area have discovered an association of tumor shape features extracted from MRI with its genomic subtypes [5,6]. However, the first step when extracting tumor features was the manual segmentation of MRI. Such annotation is costly, time consuming and results in annotations with high inter-rater variance [7].

Deep learning is a new field of machine learning that is recently revolutionizing the automated analysis of images [8,9]. There are many examples of successful applications of deep learning in medical imaging [10–14] and more specifically in brain MRI segmentation [15]. In recent years, progress in deep learning for automatic brain segmentation matured to a level that achieves performance of a skilled radiologist [16]. Most of these efforts are focused on glioblastoma rather than LGG.

\* Corresponding author.

E-mail addresses: [mateusz.buda@duke.edu](mailto:mateusz.buda@duke.edu) (M. Buda), [ashirbani.saha@duke.edu](mailto:ashirbani.saha@duke.edu) (A. Saha), [maciej.mazurowski@duke.edu](mailto:maciej.mazurowski@duke.edu) (M.A. Mazurowski).

<https://doi.org/10.1016/j.combiomed.2019.05.002>

Received 12 March 2019; Received in revised form 25 April 2019; Accepted 1 May 2019  
0010-4825/ © 2019 Published by Elsevier Ltd.

Development of models that yield high quality segmentation of LGG in brain MRI would potentially allow for automatization of the process of tumor genomic subtype identification through imaging that is fast, inexpensive, and free of inter-reader variability.

In this study, we combine the field of deep learning and radio-genomic and propose a fully automatic algorithm for quantification of tumor shape and test whether the assessed shape features are prognostic of tumor molecular subtypes. Developing imaging-biomarkers that could inform of tumor genomics would provide the information to clinicians sooner in a non-invasive way and in some cases could allow for better stratification of tumors where resection is not performed. In this study, we show promise for eventually developing such imaging-based biomarkers.

The reminder of this paper is organized as follows. Section 2. describes data used in our study whereas section 3. describes segmentation model, features used for tumor quantification, and statistical methods. Then, in section 4. we show results for the segmentation algorithm and prediction of genomic subtypes. In section 5. we discuss our findings. Finally, sections 6. and 7. are devoted to limitations and conclusions of the study, respectively.

## 2. Dataset

### 2.1. Patient population

The data used in this study was obtained from The Cancer Genome Atlas (TCGA) and The Cancer Imaging Archive (TCIA). We identified 120 patients from TCGA lower-grade glioma collection (<https://cancergenome.nih.gov/cancersselected/lowergrade glioma>) who had preoperative imaging data available, containing at least a fluid-attenuated inversion recovery (FLAIR) sequence. Ten patients had to be excluded since they did not have genomic cluster information available. The final group of 110 patients was from the following 5 institutions: Thomas Jefferson University (TCGA-CS, 16 patients), Henry Ford Hospital (TCGA-DU, 45 patients), UNC (TCGA-EZ, 1 patient), Case Western (TCGA-FG, 14 patients), Case Western – St. Joseph's (TCGA-HT, 34 patients) from TCGA LGG collection. The complete list of patients used in this study is included in Online Resource 1.

The entire set of 110 patients was split into 22 non-overlapping subsets of 5 patients each. This was done for evaluation with cross-validation.

### 2.2. Imaging data

Imaging data was obtained from The Cancer Imaging Archive (<https://wiki.cancerimagingarchive.net/display/Public/TCGA-LGG>) which contains the images corresponding to the TCGA patients and is sponsored by the National Cancer Institute. We used all modalities when available and only FLAIR in case any other modality was missing. There were 101 patients with all sequences available, 9 patients with missing post-contrast sequence, and 6 with missing pre-contrast sequence. The complete list of available sequences for each patient is included in Online Resource 1. The number of slices varied among patients from 20 to 88. In order to capture the original pattern of tumor growth, we only analyzed preoperative data. The assessment of tumor shape was based on FLAIR abnormality since enhancing tumor in LGG is rare.

A researcher in our laboratory, who was a medical school graduate with experience in neuroradiology imaging, manually annotated FLAIR images by drawing an outline of the FLAIR abnormality on each slice to form training data for the automatic segmentation algorithm. We used software developed in our laboratory for this purpose. A board eligible radiologist verified all annotations and modified those that were identified as incorrect. Dataset of registered images together with manual segmentation masks for each case used in our study is released and made publicly available at the following link: <https://www.kaggle.com/mateusz buda/lgg-mri-segmentation>.

### 2.3. Genomic data

Genomic data used in this study consisted of DNA methylation, gene expression, DNA copy number, and microRNA expression, as well as IDH mutation 1p/19q co-deletion measurement. Specifically, in our analysis we consider six previously identified molecular classifications of LGG that are known to be correlated with some tumor shape features [6]:

- 1 Molecular subtype based on IDH mutation and 1p/19q co-deletion (three subtypes: IDH mutation-1p/19q co-deletion, IDH mutation-no 1p/19q co-deletion, IDH wild type)
- 2 RNASeq clusters (4 clusters: R1-R4)
- 3 DNA methylation clusters (5 clusters: M1-M5)
- 4 DNA copy number clusters (3 clusters: C1–C3)
- 5 microRNA expression clusters (4 clusters: mi1-mi4)
- 6 Cluster of clusters (3 clusters: coc1-coc3)

## 3. Methods

### 3.1. Automatic segmentation

Fig. 1 shows the overview of the segmentation algorithm. The following phases comprise the fully automatic algorithm for obtaining the segmentation mask: image preprocessing, segmentation, and post-processing. Then, once the segmentation masks are generated we extracted shape features that were identified as predictive of molecular subtypes. The following sections provide details on each of the steps. Source code of the algorithm described in this section is also available at the following link: <https://github.com/MaciejMazurowski/brain-segmentation>.

### 3.2. Preprocessing

Images varied significantly between patients in terms of size. The preprocessing of the image sequences consisted of the following steps:

- Scaling of the images to the common frame of reference.
- Removal of the skull to focus the analysis on the brain region (a.k.a., skull stripping).
- Adaptive window and level adjustment based on the image histogram to normalize intensities of tissues between cases.
- Z-score normalization of the entire data set.

The details of all the pre-processing steps are included in the Online Resource 2.

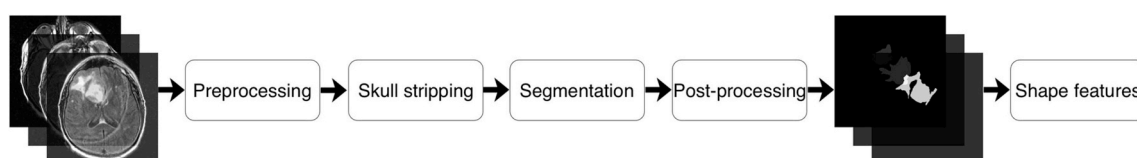


Fig. 1. A schema showing data processing steps of our system for molecular subtype inference from a sequence of brain MRI.



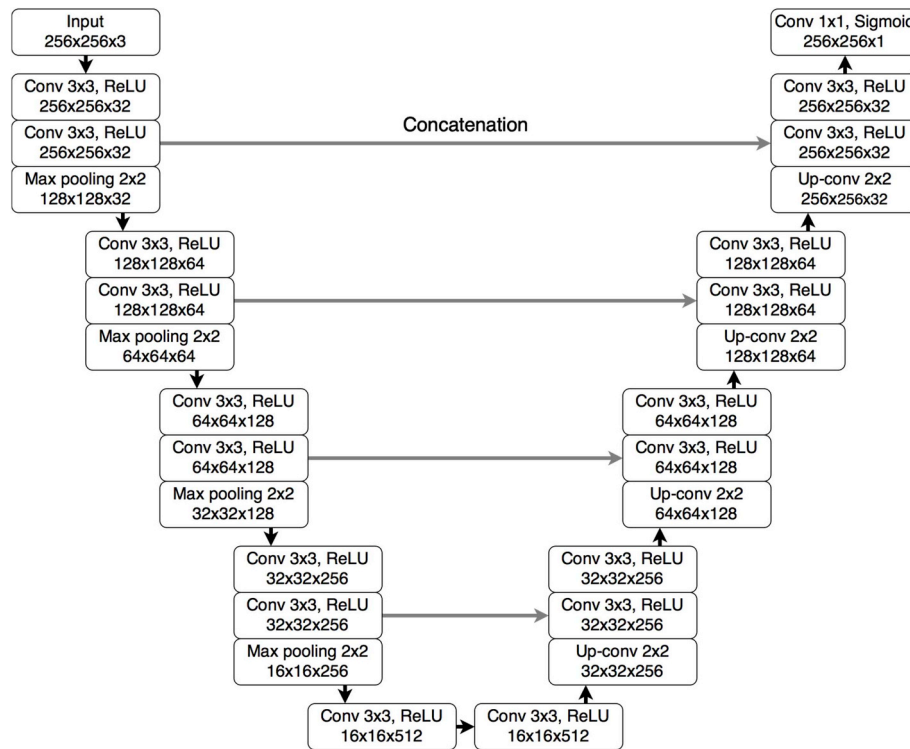


Fig. 2. U-Net architecture used for skull stripping and segmentation. Below each layer specification we provide dimensionality of a single example that this layer outputs.

### 3.3. Segmentation

The main segmentation step was performed using a fully convolutional neural network with the U-Net architecture [10] shown in Fig. 2. It comprises four levels of blocks containing two convolutional layers with ReLU activation function and one max pooling layer in the encoding part and up-convolutional layers instead in the decoding part [17–21]. Consistent with the U-Net architecture, from the encoding layers we use skip connections to the corresponding layers in the decoding part. They provide a shortcut for gradient flow in shallow layers during the training phase [22].

Manual segmentation served as a ground truth for training a model for automatic segmentation. We trained two networks, one for cases with three sequences available (pre-contrast, FLAIR, and post-contrast) and the other that used only FLAIR. For the second network, instead of missing sequences we used neighboring FLAIR slices from both sides of a slice of interest as additional channels. Since in this scenario the two sequences, which occupied channel 1 and channel 3 of the input are not available, we filled these channels with neighboring tumor slices to provide additional information to the network.

The number of slices containing tumor was considerably lower than those with only background class present. Therefore, to account for this fact, we applied oversampling with data augmentation that was proved to help in training convolutional neural networks [23]. We did it by having three instances of each tumor slice in our training set. For one oversampled slice we applied random rotation by 5–15° and for the other slice we applied random scale by 4%–8%. To further reduce the imbalance between tumor and non-tumor pixels, we discarded empty slices that did not contain any brain or other tissue after applying skull stripping. This step has been undertaken since training a fully convolutional neural network with images that do not contain any positive voxels can be highly detrimental. Please note that a significant majority of voxels in the abnormal slices are still normal and therefore sufficient negative data is available for training.

### 3.4. Post-processing

To further improve the accuracy, we implemented a post-processing algorithm that removes false positives. Specifically, we extracted all tumor volumes using connected components algorithm on a three-dimensional segmentation mask for each patient. We did it using 6-connected pixels in three dimensions, i.e. neighboring pixels are defined as being connected along primary axes. Eventually, we included in the final segmentation mask only the pixels comprising the largest connected tumor volume. This post-processing strategy benefits extraction of shape features (described in the following section) since they are sensitive to isolated false positive pixel segmentations.

### 3.5. Extraction of shape features

We consider three shape features of a segmented tumor that were identified as important in the context of lower grade glioma radio-genomic [6]:

**Angular standard deviation (ASD)** is the average of the radial distance standard deviations from the centroid of the mass across ten equiangular bins in one slice, as described in Ref. [24]. Before calculating the value of this feature, we normalize radial distances to have mean equal one. Angular standard deviation of a tumor shape is a quantitative measure of variation in the tumor margin within relatively small parts of the tumor. It also captures non-circularity of the tumor, i.e. low value indicates circle like shape.

**Bounding ellipsoid volume ratio (BEVR)** is the ratio between the volume of segmented FLAIR abnormality and its minimum bounding ellipsoid. This feature captures the irregularity of the tumor in three dimensions. If the tumor fits well into its bounding ellipsoid (high value of BEVR), it is considered more regular while if more space in the bounding ellipsoid is unfilled, the shape is considered irregular.

**Margin fluctuation (MF)** is computed as follows. First, we find the centroid of the tumor and distances from it to all pixels on the tumor boundary in one slice. Then, we apply averaging filter of length equal to

10% of the tumor perimeter measured in the number of pixels. Margin fluctuation is the standard deviation of the difference between values before and after smoothing, i.e. applying averaging filter. Similarly as in ASD, radial distances are normalized to have a mean of one. This is done in order to remove the impact of tumor size on the value of this feature. Margin fluctuation is a two-dimensional feature that quantifies the amount of high frequency changes, i.e. smoothness of the tumor boundary and was previously used for analysis of spiculation in breast tumors [25,26].

### 3.6. Statistical analysis

Our hypothesis was that fully automatically-assessed shape features are predictive of tumor molecular subtypes. Since we considered 6 definitions of molecular subtypes based on genomic assays and multiple imaging features, we focused our analysis on the relationships between imaging and genomics that were found significant (with manual tumor segmentation) in a previous study [6]. Specifically, those were the following relationships: bounding ellipsoid volume ratio with RNASeq, miRNA, CNC, and COC, the relationship of Margin fluctuation with RNASeq, and the relationship of angular standard deviation with IDH/1p19q, RNASeq, Methylation, CNC, and COC resulting in 10 specific hypotheses. To assess statistical significance of these associations, we conducted the Fisher exact test (fisher.test function in R) for each of 10 combinations of imaging and genomics. For the purpose of this test, we turned each continuous imaging variable value into a number from 1 to 4 based on which quartile of the feature value it fell into. For each of the imaging and genomic feature combinations, we used only the cases that had both: imaging and genomic subtype data available.

We conducted a total of 10 statistical tests for each pair of imaging feature and genomic subtype for our primary hypothesis. To account for multiple hypothesis testing, we applied a Bonferroni correction. *P*-values lower than 0.005 (0.05/10) were considered statistically significant for our primary radiogenomics hypotheses.

Additionally, we evaluated performance of the deep learning-based segmentation itself. We used Dice similarity coefficient [27] as the evaluation metric which measures the overlap between the segmentation provided by the algorithm and the manually-annotated gold standard.

In the evaluation process, we used cross-validation. Specifically, we divided our entire dataset into 22 subsets, each containing exactly 5 cases. The model training was conducted on the training subsets and then the model was applied to the test cases. This was repeated 22 times until each subset served once as the test set. The cases then were pooled for the analysis as described above.

The number of cases included in the training and test sets (which determines the number of folds) is a trade-off between computational cost of training multiple models and having more data to train each of them. The two extremes of this approach are leave-one-out strategy which results in one-case folds and the other is 50% split which gives 2 folds. We found folds of 5 patients to be a good balance between a training set size and computational cost.

## 4. Results

The patients' characteristics are shown in Table 1. The average patient age was 47. Fifty six of the patients were women, fifty three were men, and the gender of one was unknown. The tumors' characteristics are provided in Table 2. There were three histological types of tumors: oligodendroglioma (47), astrocytoma (33), oligoastrocytoma (29), and one unknown. Grade of the tumors in our data included 51 cases of grade II, 58 of grade III, and grade of one tumor was unknown.

The results of the radiogenomic analysis are shown in Fig. 3. We confirmed our primary hypothesis for two pairs of imaging features and genomic subtype. We found the strongest association between RNASeq cluster and the bounding ellipsoid volume ratio ( $p < 0.0002$ ) along

**Table 1**

Patient characteristic. Age for one patient was missing and was ignored in the calculation.

Characteristic	Patients (N = 110)
Age (years)	
Median	47
Range	20–75
Gender	
Female	56
Male	53
Not available	1

**Table 2**

Tumor characteristic.

Characteristic	Cases (N = 110)
Histologic type and grade	
Astrocytoma	
Grade II	8
Grade III	25
Oligoastrocytoma	
Grade II	14
Grade III	15
Oligodendroglioma	
Grade II	29
Grade III	18
Not available	1
IDH-1p/19q subtype	
IDH mutation, 1p/19q co-deletion	26
IDH mutation, no 1p/19q co-deletion	56
IDH wild type	25
Not available	3

with margin fluctuation ( $p < 0.005$ ). In addition, we identified considerable correlations for the bounding ellipsoid volume ratio and all tested molecular subtypes ( $p < 0.02$ ) as well as for angular standard deviation and RNASeq cluster ( $p < 0.02$ ).

In Table 3 we provide ROC AUC scores for the task of discriminating each RNASeq cluster from all other subtypes based on shape features extracted from segmentation masks obtained with the deep learning algorithm and compare them to manual segmentations. We selected RNASeq cluster since it showed the strongest association with shape features. In addition, we included ROC AUC based on two demographic variables, i.e. age and gender. The results show that deep learning was able to provide tumor segmentations of a quality that allowed for extraction of shape features that match manual segmentations. Specifically, cluster R2 was distinguished from all other clusters based on inversed bounding ellipsoid volume ratio with AUC of 0.80 and 0.78 for deep learning-based and manual segmentations, respectively. In terms of angular standard deviation, AUC for deep learning was 0.73 and for manual segmentations was 0.72. The predictive value of demographic variables for R2 cluster was notably lower with AUC = 0.66 for age and AUC = 0.50 for gender. A detailed comparison of manual and automatic segmentation for the task of discriminating cluster R2 from all other clusters with respect to sensitivity, specificity, positive predictive value, and negative predictive value is given in Online Resource 1.

In terms of tumor segmentation, our deep learning algorithm achieved mean Dice coefficient of 82% and median Dice coefficient of 85%. For the 101 cases with all sequences available, mean and median Dice coefficient was the same as for all 110 cases. For the remaining 9 cases segmented based on FLAIR sequence only, mean and median Dice coefficient was 82% and 88%, respectively. Examples of segmentations that we obtained alongside with ground truth masks for cases of varying performance of our segmentation algorithm are presented in Fig. 4. Due to max pooling layers included in the U-Net architecture which allow for processing of large volumes on currently available computers, the automated segmentation is less sensitive to high curvature and

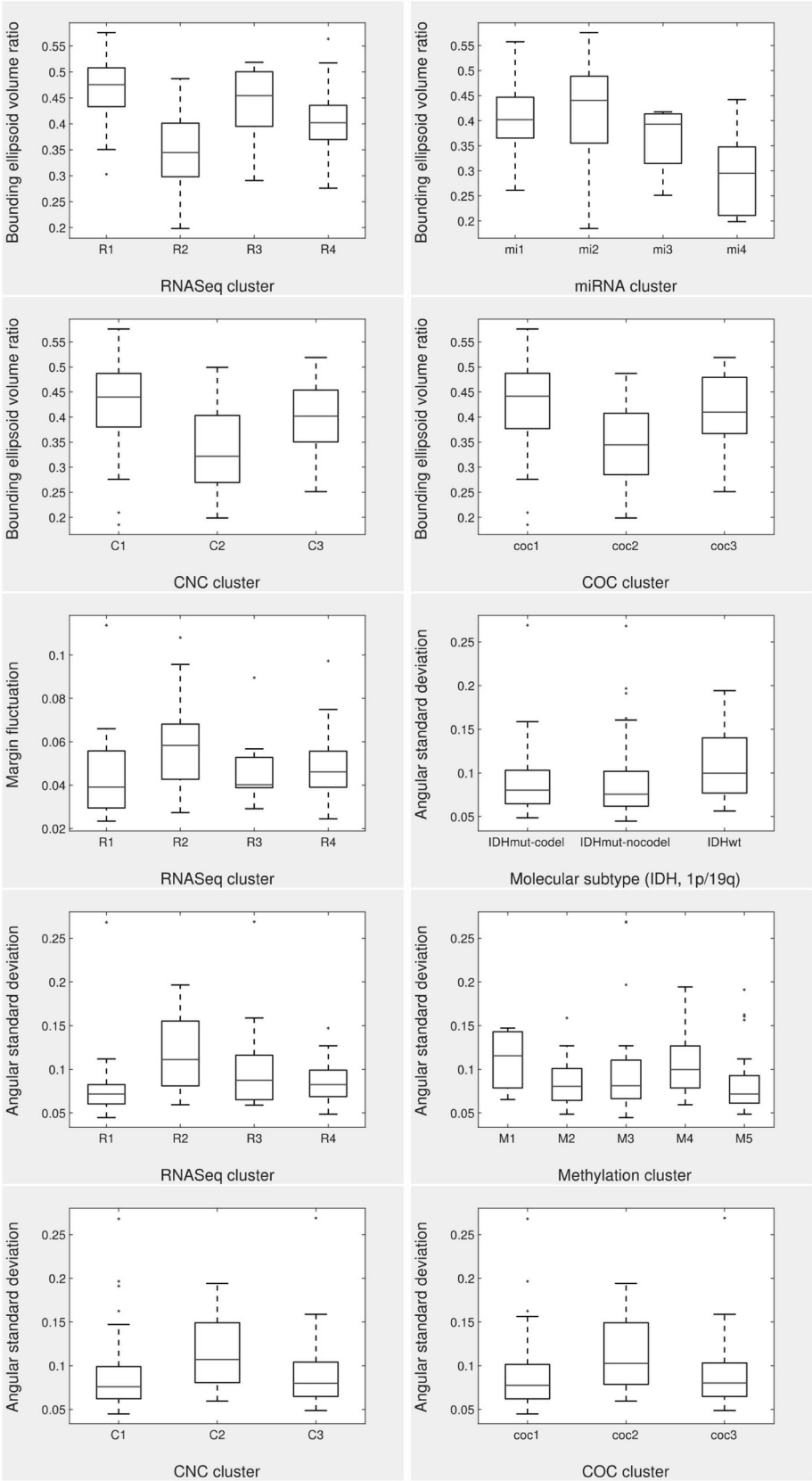


Fig. 3. Box and whisker plots demonstrating the relationship between tested genomic clusters and imaging features that quantify tumor shape.

**Table 3**  
ROC AUCs of shape features and demographic variables for the task of discriminating one RNASeq cluster from all others.

Variable	R1	R2	R3	R4
Deep learning ASD	0.26	0.73	0.53	0.47
Deep learning 1/BEVR	0.23	0.80	0.36	0.53
Manual ASD	0.17	0.72	0.48	0.60
Manual 1/BEVR	0.21	0.78	0.36	0.57
Age	0.29	0.66	0.72	0.41
Gender	0.44	0.50	0.58	0.52

sulcation and therefore, the produced masks tend to be more smooth comparing to manual segmentations.

Given small sample size, we additionally assessed the stability and performance of our segmentation model using predictions generated for noisy input. For each input image, we applied additive gaussian noise with zero mean and standard deviation equal to 10% and 20% of standard deviation computed on the training data. In the first case with 10% noise level, mean Dice coefficient decreased to 81%. After increasing noise level to 20%, mean Dice coefficient further decreased to 79%. In both cases, median Dice coefficient was 85%.

5. Discussion

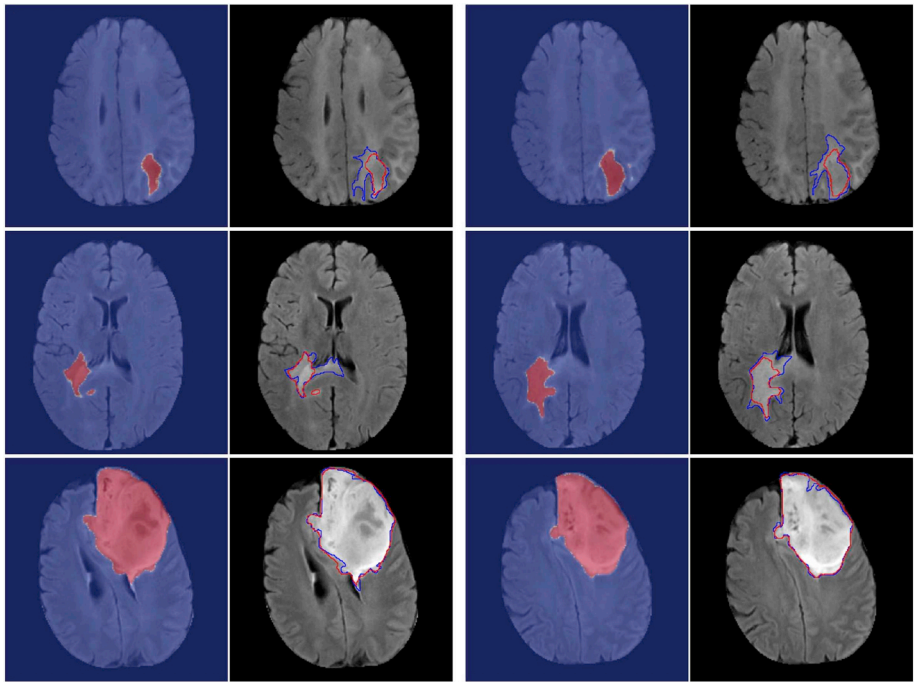
In this study, we were able to demonstrate that fully automatically-assessed imaging features of low grade gliomas are associated with tumor molecular subtypes established using genomic analysis. The strength of these associations was shown to be moderate. Deep learning algorithms were used to segment the tumors.

Using imaging to predict tumor genomics is of very high importance and if accurate models are developed, it could be incorporated in the current treatment paradigm in a variety of ways. In the simplest scenario, if the model is highly accurate, it could simply replace the genomic analysis altogether. Current state of the art radiogenomic models, such as the one shown in this paper which represents moderate predictive performance, do not yet justify such substitution. However, there are other ways in which even models with moderate performance could contribute valuable information. The imaging data is available early in the process and therefore approximate assessment of tumor

biology before the surgery could still be of help in guiding the next steps. The approximate imaging surrogate of molecular subtype would also be of particular use for patients that do not immediately undergo surgical excision of the tumor. In such case, in the absence of tissue analysis, the approximate classification by imaging could be of very high value since genomic subtypes are highly correlated with patient outcomes. If biopsy results are available for a tumor that has not been fully resected, the imaging surrogate of subtype can still be of use given a potential high intra-tumor heterogeneity of the lesions and therefore a possibility that a local biopsy does not accurately reflect the overall genomics of a tumor. Imaging offers a complete view of a tumor. Finally, even if the overall accuracy is not perfect, it might be possible to operate at a high positive predictive value or high predictive value and the surrogate imaging-based models of genomic subtypes could be still used for triaging patients for genomic tests, even if it is a small minority of the patients. For example, if based on imaging there is a high confidence that a tumor is of a particular aggressive subtype, the patient could be treated accordingly without additional expensive and invasive genomic testing. If on the other hand the imaging-based marker has low confidence, then additional genomic test could be ordered.

An important step toward accurate and reproducible assessment of imaging features of lower-grade gliomas is accurate segmentation of the tumors. While the annotation by radiologists is considered a gold standard, a considerable inter-observer variability has been documented for this task. For the whole tumor segmentation of LGG on brain MRI, Dice coefficient between two expert raters is 84% with standard deviation of 2% [7]. This demonstrates that our algorithm falls within acceptable level precision. At the same time, our algorithm provides a fully reproducible and consistent way of tumor quantification for future cases.

Automatic segmentation of tumors such as the one showed in this study has multiple advantages. First, it addresses the inter-observer variability described above. Since there is only one reader (the computer algorithm), the inter-observer variability is non-existent. Furthermore, it addresses the problem of intra-observer variability. The algorithm is deterministic which means that given the same image, the algorithm will always perform an identical assessment. Finally, application of a computer algorithm is inexpensive and fast. The



**Fig. 4.** Examples of automatic segmentation for low (top), moderate (middle), and high (bottom) agreement with ground truth. Their Dice coefficient is 50%, 82% and 95% respectively. In each pair, the first image shows a heatmap of raw model output and in the second image blue outline corresponds to ground truth and red to postprocessed automatic segmentation output. Images show FLAIR modality after preprocessing and skull stripping. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)



performance of our segmentation algorithm in terms of the mean Dice coefficient was 82% which puts it on a par with expert human readers. This was achieved with the help of deep learning which has demonstrated similar phenomenal performance in other applications.

Our results show that RNASeq R2 cluster, compared to other clusters, is associated with the tumors of notably higher irregularity of shape as quantified by bounding ellipsoid volume ratio, angular standard deviation and margin fluctuation. R2 cluster is in turn linked to considerably poorer overall survival as compared with R1, R3, and R4 [1]. The same conclusion can be drawn for molecular subtype of IDH wild type which indicates less favorable prognosis that is close to glioblastoma prognosis [1]. It is associated with relatively high angular standard deviation. This is consistent with findings obtained in the previous study for shape features extracted from manually segmented tumors [6]. This points out to a conclusion that angular standard deviation, margin fluctuation, bounding ellipsoid volume ratio, and potentially other features that measure the irregularity of tumor shape may be prognostic of patient's outcome.

## 6. Limitations

This study had limitations. It constitutes only a first step toward imaging-based surrogates of genomic subtypes. Specifically, only three imaging features were considered. While these features were selected based on prior evidence of their effectiveness and therefore are of high importance, they constitute a small sample of different features that could be calculated including texture and enhancement of the tumor and its surroundings. Furthermore, a fairly limited sample size was used in the study (110 patients) since data that contains comprehensive genome-wide assays alongside with imaging is still rare. While no separate validation set was available in this study, we utilized a commonly used cross-validation technique, which splits the data into training and test sets to avoid a positive evaluation bias.

Regarding segmentation algorithms, there are many methods for performing automatic segmentation of brain tumors that could be considered for comparison and to further improve our results [28,29]. First, regarding general approaches to segmentation, in Ref. [16] deep learning with sliding-window approach was used. An improvement that significantly reduces computational complexity is a fully convolutional neural network which allows for processing entire image in one forward pass [30,31]. In U-Net architecture, used in our study, additional skip connections between encoder and decoder parts of the network are used [10]. Second, regarding network architecture, different types have been proposed, e.g. ResNet [32], Inception [33], and DenseNet [34], which were incorporated in segmentation models. Finally, various optimization functions for training deep learning segmentation models were proposed. The most commonly applied is cross-entropy loss, used also in classification models. However, for highly imbalanced segmentation tasks, loss functions based on Dice similarity coefficient outperformed other loss functions in many applications [22,35–37].

## 7. Conclusions

In conclusion, we demonstrated that features of MRI, extracted in a fully automatic manner using deep learning algorithms, were associated with tumor molecular subtypes of lower-grade gliomas determined using genomic assays. This shows promise for reproducible non-invasive imaging-based surrogates of tumor genomics in brain cancer.

## Informed consent

This article does not contain any studies with human participants or animals performed by any of the authors.

## Conflicts of interest

All authors declare that they have no conflict of interest.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.compbimed.2019.05.002>.

## References

- [1] C.G.A.R. Network, others, Comprehensive, Integrative genomic analysis of diffuse lower-grade gliomas, *N. Engl. J. Med.* 2015 (2015) 2481–2498.
- [2] C.-M. Zhang, D.J. Brat, Genomic profiling of lower-grade gliomas uncovers cohesive disease groups: implications for diagnosis and treatment, *Chin. J. Canc.* 35 (2016) 12.
- [3] J.E. Eckel-Passow, D.H. Lachance, A.M. Molinaro, K.M. Walsh, P.A. Decker, H. Sicotte, M. Pekmezci, T. Rice, M.L. Kosel, I.V. Smirnov, Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors, others, *N. Engl. J. Med.* 372 (2015) 2499–2508.
- [4] M.A. Mazurowski, Radiogenomics: what it is and why it is important, *J. Am. Coll. Radiol.* 12 (2015) 862–866.
- [5] M.A. Mazurowski, K. Clark, N.M. Czarnek, P. Shamsesfandabadi, K.B. Peters, A. Saha, Radiogenomic analysis of lower grade glioma: a pilot multi-institutional study shows an association between quantitative image features and tumor genomics, *Proc. SPIE* (2017) vol. 10134, doi: 10.1117/12.2255579.
- [6] M.A. Mazurowski, K. Clark, N.M. Czarnek, P. Shamsesfandabadi, K.B. Peters, A. Saha, Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the Cancer Genome Atlas data, *J. Neuro Oncol.* (2017) 1–9.
- [7] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, others, the multimodal brain tumor image segmentation benchmark (BRATS), *IEEE Trans. Med. Imaging* 34 (2015) 1993–2024.
- [8] I. Goodfellow, Deep Learning of Representations and its Application to Computer Vision, (2015).
- [9] A. Ioannidou, E. Chatzilaris, S. Nikolopoulos, I. Kompatsiaris, Deep learning advances in computer vision with 3D data: a survey, *ACM Comput. Surv.* 50 (2017) 20.
- [10] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, *Int. Conf. Med. Image Comput. Comput. Interv.* 2015, pp. 234–241.
- [11] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, *Int. Conf. Med. Image Comput. Comput. Interv.* 2016, pp. 424–432.
- [12] F. Milletari, N. Navab, S.-A. Ahmadi, V-net, Fully convolutional neural networks for volumetric medical image segmentation, *3D Vis. (3DV)*, 2016 Fourth Int. Conf., 2016, pp. 565–571.
- [13] H.P.-A. Chen Hao, Qi Dou, Lequan Yu, Jing Qin, VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images, *Neuroimage* 170 (2018) 446–455.
- [14] M.A. Mazurowski, M. Buda, A. Saha, M.R. Bashir, Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI, *J. Magn. Reson. Imaging* (2019), 939–954.
- [15] M. Havaei, N. Guizard, H. Larochelle, P.-M. Jodoin, Deep learning trends for focal brain pathology segmentation in MRI, *Mach. Learn. Heal. Informatics*, Springer, 2016, pp. 125–148.
- [16] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, H. Larochelle, Brain tumor segmentation with deep neural networks, *Med. Image Anal.* 35 (2017) 18–31.
- [17] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, *Proc. Fourteenth Int. Conf. Artif. Intell. Stat.* 2011, pp. 315–323.
- [18] Y. LeCun, Y. Bengio, others, Convolutional networks for images, speech, and time series, *Handb. Brain Theory Neural Networks*, vol. 3361, 1995, p. 1995.
- [19] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE*, 1998.
- [20] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, *Adv. Neural Inf. Process. Syst.* 2012, pp. 1097–1105.
- [21] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, *Proc. IEEE Int. Conf. Comput. Vis.* 2015, pp. 1520–1528.
- [22] M. Drozdal, E. Vorontsov, G. Chartrand, S. Kadoury, C. Pal, The importance of skip connections in biomedical image segmentation, *Int. Work. Large-Scale Annot. Biomed. Data Expert Label Synth.* 2016, pp. 179–187.
- [23] M. Buda, A. Maki, M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, *Neural Network*. 106 (2018) 249–259, <https://doi.org/10.1016/J.NEUNET.2018.07.011>.
- [24] H. Georgiou, M. Mavroforakis, N. Dimitropoulos, D. Cavouras, S. Theodoridis, Multi-scaled morphological features for the characterization of mammographic masses using statistical classification schemes, *Artif. Intell. Med.* 41 (2007) 39–55.
- [25] M.L. Giger, C.J. Vyborny, R.A. Schmidt, Computerized characterization of mammographic masses: analysis of spiculation, *Cancer Lett.* 77 (1994) 201–211.
- [26] S. Pohlman, K.A. Powell, N.A. Obuchowski, W.A. Chilcote, S. Grundfest-

- Broniatowski, Quantitative classification of breast tumors in digitized mammo-grams, *Med. Phys.* 23 (1996) 1337–1345.
- [27] L.R. Dice, Measures of the amount of ecologic association between species, *Ecology* 26 (1945) 297–302.
- [28] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, *Med. Image Anal.* 42 (2017) 60–88, <https://doi.org/10.1016/j.media.2017.07.005>.
- [29] Z. Akkus, A. Galimzianova, A. Hoogi, D.L. Rubin, B.J. Erickson, Deep learning for brain MRI segmentation: state of the art and future directions, *J. Digit. Imaging* 30 (2017) 449–459.
- [30] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, (2015) ArXiv Prepr. ArXiv1511.00561.
- [31] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
- [32] Q. Zhang, Z. Cui, X. Niu, S. Geng, Y. Qiao, Image segmentation with pyramid dilated convolution based on ResNet and U-net, *Int. Conf. Neural Inf. Process.*, 2017, pp. 364–372.
- [33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
- [34] S. Jégou, M. Drozdal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation, *Comput. Vis. Pattern Recognit. Work. (CVPRW)*, 2017 IEEE Conf., 2017, pp. 1175–1183.
- [35] C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M.J. Cardoso, Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations, *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, Springer, 2017, pp. 240–248.
- [36] L. Fidon, W. Li, L.C. Garcia-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin, T. Vercauteren, Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks, *Int. MICCAI Brainlesion Work.*, 2017, pp. 64–76.
- [37] J. Zhang, X. Shen, T. Zhuo, H. Zhou, Brain Tumor Segmentation Based on Refined Fully Convolutional Neural Networks with a Hierarchical Dice Loss, (2017) ArXiv Prepr. ArXiv1712.09093.

#### **A.4 Deep radiogenomics of lower-grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images**

# Deep Radiogenomics of Lower-Grade Gliomas: Convolutional Neural Networks Predict Tumor Genomic Subtypes Using MR Images

Mateusz Buda, MSc • Ehab A. AlBadawy, BSc • Ashirbani Saha, PhD • Maciej A. Mazurowski, PhD

From the Department of Radiology, Duke University School of Medicine, 2424 Erwin Rd, Suite 302, Durham, NC 27705 (M.B., E.A.A., A.S., M.A.M.); Department of Electrical and Computer Engineering, Duke University, Durham, NC (M.A.M.); and Department of Biostatistics and Bioinformatics, Duke University, Durham, NC (M.A.M.). Received October 9, 2018; revision requested November 23, 2018; revision received August 6, 2019; accepted August 30, 2019. Address correspondence to M.B. (e-mail: [mateusz.buda@duke.edu](mailto:mateusz.buda@duke.edu)).

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2020; 2(1):e180050 • <https://doi.org/10.1148/ryai.2019180050> • Content codes:   

**Purpose:** To employ deep learning to predict genomic subtypes of lower-grade glioma (LLG) tumors based on their appearance at MRI.

**Materials and Methods:** Imaging data from The Cancer Imaging Archive and genomic data from The Cancer Genome Atlas from 110 patients from five institutions with lower-grade gliomas (World Health Organization grade II and III) were used in this study. A convolutional neural network was trained to predict tumor genomic subtype based on the MRI of the tumor. Two different deep learning approaches were tested: training from random initialization and transfer learning. Deep learning models were pretrained on glioblastoma MRI, instead of natural images, to determine if performance was improved for the detection of LGGs. The models were evaluated using area under the receiver operating characteristic curve (AUC) with cross-validation. Imaging data and annotations used in this study are publicly available.

**Results:** The best performing model was based on transfer learning from glioblastoma MRI. It achieved AUC of 0.730 (95% confidence interval [CI]: 0.605, 0.844) for discriminating cluster-of-clusters 2 from others. For the same task, a network trained from scratch achieved an AUC of 0.680 (95% CI: 0.538, 0.811), whereas a model pretrained on natural images achieved an AUC of 0.640 (95% CI: 0.521, 0.763).

**Conclusion:** These findings show the potential of utilizing deep learning to identify relationships between cancer imaging and cancer genomics in LGGs. However, more accurate models are needed to justify clinical use of such tools, which might be obtained using substantially larger training datasets.

*Supplemental material is available for this article.*

© RSNA, 2020

Lower-grade gliomas (LGGs) are a diverse group of brain tumors classified as grade II and III using the World Health Organization grading system. Histopathologic analysis suffers from interobserver variability and can be inaccurate in predicting patient outcomes (1). Recently, a new tumor subtyping scheme was proposed which clusters LGGs based on DNA methylation, gene expression, DNA copy number, and microRNA expression (1). In particular, unsupervised analysis of tumors based on their molecular profiles derived from these four platforms resulted in a second-level cluster of clusters (CoC) partitioning into three distinctive biologic subsets (CoC1 to CoC3). It has been shown that the new subtypes are, to a large extent, in agreement with more basic subtyping utilizing isocitrate dehydrogenase (*IDH1* and *IDH2*) mutation and 1p/19q codeletion (1,2). It has been determined that tumors from the different molecular groups substantially differ in terms of typical course of the disease and overall survival (3). Specifically, the CoC2 cluster was found to have a strong correlation with wild-type *IDH* molecular subtype and had an overall survival rate similar to that of glioblastoma (GBM).

Radiogenomics is a new direction in cancer research that aims to identify relationships between tumor genomic

characteristics and imaging phenotypes (ie, its presentation on radiologic images) (4). In addition to extending our understanding of the disease in general, radiogenomics might provide actionable information if the genomic characteristics of tumors can be predicted prior to invasive tissue examination or in cases when resection is risky or impossible. Some radiogenomic studies of LGGs have discovered an association of tumor shape features extracted from brain MRI with its genomic subtypes (5–7). A shortcoming of the previously proposed method is that the features of the image used for the analysis need to be decided a priori without knowing which image characteristics might be most predictive of tumor genomics. Often a very large number of features are extracted, which increases the likelihood of the noisy features obscuring the important ones. An alternative, more holistic approach, proposed in this study, is based on a supervised deep learning model that allows the algorithm to learn which imaging characteristics are the most helpful in making the prediction.

In recent years, progress in deep learning has allowed for the development of highly accurate models for various image-related tasks, even in the presence of limited training data (8,9). Although neural networks can be trained



## Abbreviations

AUC = area under the receiver operating characteristic curve, CoC = cluster of clusters, CI = confidence interval, CNN = convolutional neural network, FLAIR = fluid-attenuated inversion recovery, GBM = glioblastoma, IDH = isocitrate dehydrogenase, LGG = lower-grade glioma, TCGA = The Cancer Genome Atlas, TCIA = The Cancer Imaging Archive

## Summary

Deep learning algorithms, especially those utilizing transfer learning, were able to find the association between imaging and genomics of lower grade gliomas.

## Key Points

- While deep learning cannot yet replace genomic testing, it shows promise in aiding clinical decisions of lower grade gliomas.
- A convolutional neural network pretrained with brain MRI of glioblastoma tumors achieved the best performance as compared with networks trained from scratch or pretrained on natural images.
- For discriminating cluster-of-clusters 2 from others, we achieved area under the receiver operating characteristic curve of 0.730 (95% confidence interval: 0.605, 0.844).

from a random initialization of the weights, an approach that has shown promise is transfer learning, which allows for pre-training of the network with a dataset different than the main training set. The performance of these methods depends on the task at hand, available data, and potentially other factors. Particularly, the type of dataset used for pretraining could be a factor influencing the final performance of the network. In this study, we tested whether a deep learning model with transfer learning from GBM MRI, instead of natural images, can provide improvement in performance over a model trained from scratch for predicting CoC molecular subtype based on MRI of LGG. The hypothesis was that a supervised deep learning approach based on MR images of LGGs would be predictive of the tumor genomics. We also determined whether a model pretrained on another MRI dataset showed better performance and generalization ability than other learning approaches.

## Materials and Methods

### Patient Population

In this institutional review board–exempt study, we analyzed patient data from The Cancer Genome Atlas (TCGA) LGG collection (10). First, we excluded patients who did not have preoperative fluid-attenuated inversion recovery (FLAIR) imaging data available. From the 120 cases that remained, we further excluded 10 patients for whom the relevant genomic data were not available. The final analyzed cohort of 110 patients contained data from the following five institutions: Thomas Jefferson University (TCGA-CS, 16 patients), Henry Ford Hospital (TCGA-DU, 45 patients), University of North Carolina (TCGA-EZ, one patient), Case Western (TCGA-FG, 14 patients), Case Western–St. Joseph's (TCGA-HT, 34 patients). These 110 patients were used for the development of classification networks and in the radiogenomic analysis. The full list of patients included in our analysis is available in Appendix E1 (supplement).

**Table 1: Imaging Data for 110 Patients Used in this Study**

Parameter	Value
No. of patients	110
No. of slices per patient*	35.7 ± 15.2
No. of tumor slices per patient*	12.3 ± 6.2
Image width†	192–512
Image height†	224–512

\* Data are means ± standard deviations.

† Data are minimum to maximum pixel measurement.

### Imaging Data

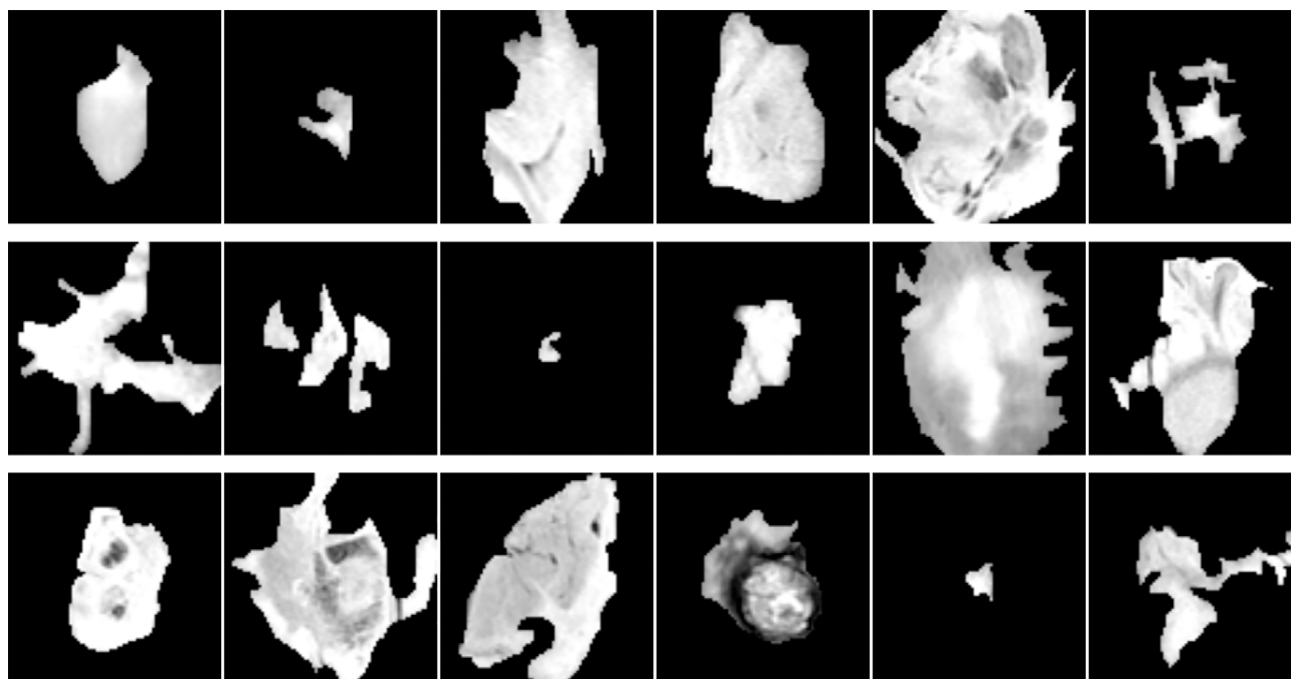
We obtained imaging data for our study from The Cancer Imaging Archive (TCIA) sponsored by the National Cancer Institute, which contains MR images for some of the patients from TCGA repository. All 110 patients included in our analysis had FLAIR sequences available, which we used for prediction of molecular subtypes. The number of slices in each sequence varied from 20 to 88. The size of images ranged from 256 × 192 to 512 × 512 pixels. Voxel spatial dimensions ranged from 0.39 to 1.02 mm and slice thickness was between 2 and 7.5 mm. All images were saved in 8-bit gray-scale lossless tagged image file format (TIFF). The characteristics of the imaging data are summarized in Table 1. Additional imaging metadata for each patient is provided in Appendix E1 (supplement). The FLAIR abnormality for tumors on all images was manually outlined by a researcher in our laboratory who used an in-house MATLAB tool developed for this task. The final annotations were approved by a board-eligible radiologist. All imaging data (preprocessed and labeled images from the TCGA-LGG collection) and annotations used in this study were made available at the following link: <https://www.kaggle.com/mateuszbudal/lgg-mri-segmentation>.

### Genomic Data

We used genomic classifications developed in a recent publication (1) defining the molecular landscape of LGGs. Genomic data came from TCGA-LGG collection and were derived based on DNA methylation, gene expression, DNA copy number, microRNA expression, and the measurement of *IDH* mutation and 1p/19q codeletion. Specifically, in our analysis we considered three CoC molecular subtypes: CoC1, CoC2, and CoC3. This subclassification has shown a strong correlation with imaging data using handcrafted tumor shape features in a previous study (6). Our data contained 55 cases for cluster CoC1, 25 cases for CoC2, and 30 cases for CoC3.

### Deep Learning for Prediction of Molecular Subtypes Based on MR Images

**Preparation of the data for training of neural networks.**—To obtain comparable results between all tested deep learning methods, we applied the same preprocessing of images across different methods. In a series of preliminary experiments, we



**Figure 1:** Example patches for tumors from cluster CoC1 (top), cluster CoC2 (middle), and cluster CoC3 (bottom) before applying rotation and shift augmentation. CoC = cluster of clusters.

identified the following transformations and data preparation steps to be essential to achieve satisfactory results for classification of genomic subtypes. All slices were first padded to square aspect ratio, resized to  $256 \times 256$  pixels, and were contrast-normalized by stretching pixel values between 1st and 99th percentile in the histogram. Then, we applied a mask from manual segmentation of tumors to guide network and provide shape information. Finally, image patches used for training and inference were cropped to  $80 \times 80$  pixels centered in the middle of the tumor. Only the slices that contained some tumor were considered. The optimal patch size was chosen based on a series of preliminary experiments.

The total number of extracted patches was 1648. Example patches for each cluster are shown in Figure 1. In addition, we performed data augmentation to generate extra training examples, a common technique in deep learning (11). Specifically, each patch was repeated five times with random rotation by  $\pm 10$  degrees and random shift by  $\pm 16$  pixels in horizontal and vertical direction, then sampled independently. This procedure resulted in 8240 examples in total. To alleviate the problem of imbalance, we applied random minority oversampling to make class distribution uniform (12).

**Training custom network from random initialization.**—The first tested approach was training a custom network from random initialization of weights (aka from scratch). The architecture of our trained-from-scratch network consisted of three standard blocks with convolutional layer, rectified linear unit, activation, max-pooling layer, and batch normalization (13–16). After that, we added two fully connected layers followed by dropout layers with 50% dropout ratio (17). The last layer contained three output units corresponding to predicted clus-

ters. A detailed description of the architecture and the training hyperparameters are provided in Appendix E2 (supplement).

**Transfer learning.**—It has been shown that deep convolutional neural networks (CNNs) trained on large datasets learn general feature representations (18,19). Shallow filters detect simple shapes (eg, edges) whereas deeper layers are responsible for recognizing more complex structures and objects (20). The most common transfer learning method is fine-tuning of a model trained on another dataset. It involves training the final classification layer from random initialization and adjustment of weights in early layers using a small learning rate. In our experiments, we fine-tuned GoogLeNet (21) network pretrained on ImageNet dataset of natural images (22). We also fine-tuned a CNN developed for classification of patches extracted from another brain MRI dataset of patients with GBM. The network was trained to distinguish different parts of the tumor and normal brain tissue with the ultimate goal of segmenting the images (23). For both of the fine-tuned models, the fully connected layers were replaced and randomly initialized. To match the number of input channels, we repeated FLAIR sequence three times. Detailed description of the network pretrained on GBM MRI data and the training hyperparameters are provided in Appendix E2 (supplement).

### Model Evaluation and Statistical Analysis

We performed the evaluation using 22-fold cross-validation. Specifically, we split the data by patients into 22 folds with five cases each. Then we trained the model using 21 folds (105 cases) and tested it using one fold (five cases). We repeated the process 22 times such that each fold was used as the test set once. Because each patient had several slices con-

**Table 2: Patient and Tumor Characteristic**

Characteristic	Age and No. of Patients ( <i>n</i> = 110)
Age (y)	
Median	47
Range	20–75
Sex	
Female	56
Male	53
Not available	1
Histologic type and grade	
Astrocytoma	
Grade II	8
Grade III	25
Oligoastrocytoma	
Grade II	14
Grade III	15
Oligodendroglioma	
Grade II	29
Grade III	18
Not available	1

Note.—Age for one patient was missing and was ignored in the calculation.

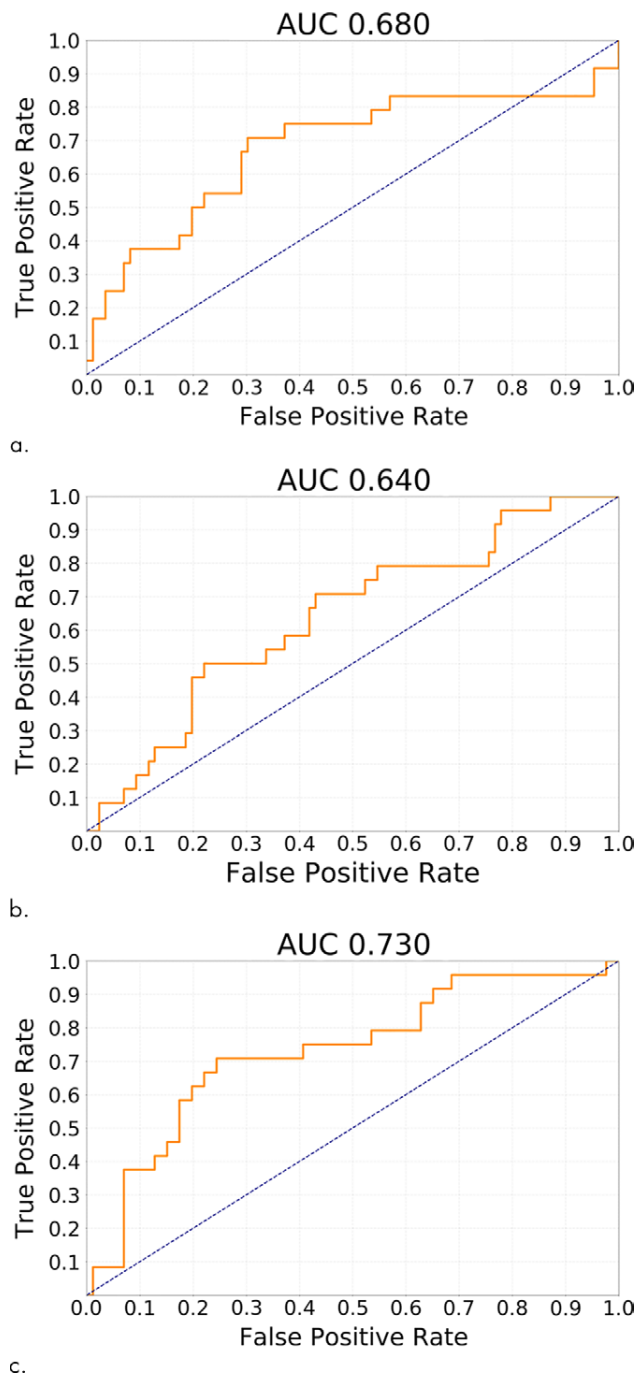
taining tumor and we trained classifiers to predict molecular subtype of a single image, we averaged the predicted scores across tumor slices, independently for each class, to arrive at the final prediction.

We used the area under the receiver operating characteristic curve (AUC) (24) computed by pooling predictions from all folds, as the evaluation metric. We evaluated how well the classifier can distinguish each given subtype (CoC1, CoC2, CoC3) from all other subtypes combined (eg, CoC1 vs CoC2 and CoC3) as well as all possible pairs for clusters (ie, CoC1 vs CoC2, CoC1 vs CoC3, CoC2 vs CoC3). For evaluation of all these binary tasks we trained a single multiclass neural network with three outputs corresponding to probabilities of three CoC clusters. In each case, we took the score from the CNN for a given class, averaged across slices as the score for computing receiver operating characteristic curves. Our particular focus was on cluster CoC2 which has been shown to be associated with a lower survival (1). Statistical tests for comparison of models and computation of confidence intervals (CIs) was performed using a bootstrapping tool implemented in Python.

## Results

The characteristics of our patient population are shown in Table 2. The average age was 47 years (one unknown). Fifty-six patients were women and 53 were men (one unknown). Among 109 patients with histologic data present, 47 were oligodendrogliomas, 29 were oligoastrocytomas, and 33 were astrocytomas. In terms of tumor grade, 51 tumors were of grade II and 58 were of grade III.

The results of testing our methods for the task of discriminating cluster CoC2 from all other clusters in terms of receiver



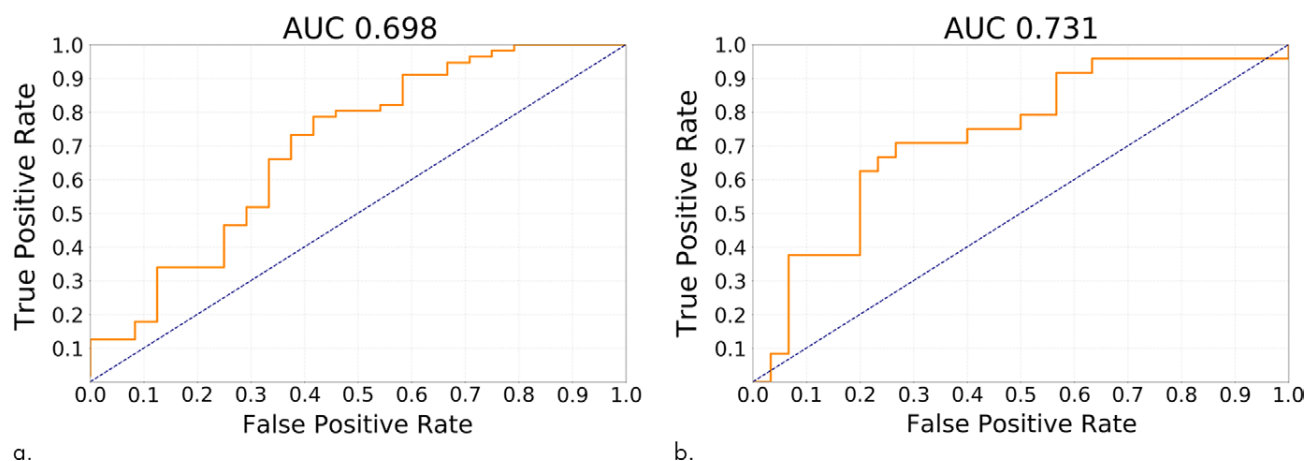
**Figure 2:** Receiver operating characteristic curves for the task of discriminating cluster CoC2 from all other clusters (CoC1 and CoC3) combined for (a) training from scratch, (b) transfer learning from ImageNet, and (c) transfer learning from glioblastoma MRI. AUC = area under the receiver operating characteristic curve, CoC = cluster of clusters.

operating characteristic curves are presented in Figure 2. The best performing method was transfer learning utilizing GBM MRI for pretraining with AUC of 0.730 (95% CI: 0.605, 0.844). In comparison, for the network trained from scratch, AUC was 0.680 (95% CI: 0.538, 0.811) and for GoogLeNet pretrained on natural images, it was 0.640 (95% CI: 0.521, 0.763). The

**Table 3: AUC with 95% Confidence Intervals for the Transfer Learning from GBM MRI Experiment**

Output Cluster	2	3	One versus All
1	0.698 (0.554, 0.823)	0.613 (0.486, 0.736)	0.650 (0.540, 0.751)
2	...	0.731 (0.591, 0.859)	0.730 (0.605, 0.844)
3	...	...	0.584 (0.449, 0.710)

Note.—AUC = area under the receiver operating characteristic curve, GBM = glioblastoma.



**Figure 3:** Receiver operating characteristic curves for transfer learning from glioblastoma MRI experiment for the task of discriminating (a) cluster CoC1 versus CoC2 and (b) cluster CoC2 versus CoC3. AUC = area under the receiver operating characteristic curve, CoC = cluster of clusters.

differences between GBM pretrained model and other models were not statistically significant ( $P > .1$ ). All deep learning methods showed performance statistically significantly higher than chance (ie, none of the CIs overlap with AUC = 0.5).

For the transfer learning method using GBM data for pretraining, Table 3 shows the ability of the deep learning method to classify different subtypes. The classifier showed the highest predictive ability for distinguishing between CoC2 and CoC3 and the lowest for distinguishing CoC1 and CoC3. Figure 3 offers a visual representation of these results. In Figure 4, we show network attention heatmaps, which indicate parts of the image responsible for prediction. Increased response by the network was for tumor margin regions of high irregularity which provides additional validation of results from previous studies (6). Additional results for discriminating between all possible combinations of CoC clusters for the two other deep learning methods tested in the study are provided in Appendix E3 (supplement).

## Discussion

In this study, we demonstrated that deep learning–based algorithms are capable of classifying molecular subtypes of LGG tumors with a moderate performance. The model that showed the highest AUC utilized previous GBM imaging data for model pretraining.

Although at this stage of the development, the imaging-based models could not be used as a one-to-one replacement for genomic testing, the correlations between genomics and imaging data are important to identify and can be applied in various ways. First, the genomic assays described in this article

are very expensive and are rarely acquired in the clinical setting. Therefore, even an imprecise prediction of a sophisticated genomics subtype could be of value in deciding the course of treatment. Second, even if a sophisticated genomic analysis is planned, it requires extraction of tumor tissue and additional time for analysis. This step means that genomic information is delayed, particularly for patients who do not undergo immediate surgery. The approximate information provided by already available imaging could immediately help with the decision process during the time when genomic information is absent. Third, an imperfect, but sufficiently accurate, model could stratify patients for genomic testing and limit the testing only to the patients where the imaging-based surrogate is not confident about the prediction. Finally, in addition to the potential clinical uses just described, the ability of deep learning to identify some characteristics of images that represent the underlying genomics could be of high value in further understanding of genotype-phenotype relationships in cancer.

The imaging-based approach to identifying the underlying tumor genomics has some very clear strengths. In addition to the low cost (MRI is already available) and immediate access to the information, imaging offers a way to analyze the tumor as a whole rather than individual tissue samples. This allows for visualizing the tumor in its surrounding and the ability to assess tumor shape, which reflects the growth pattern as well as tumor enhancement which illustrates its vascular structure. Finally, the overall look at a tumor is of utmost importance given the intratumor genomic heterogeneity of cancer. While the results of a genomic test can differ based on which part of



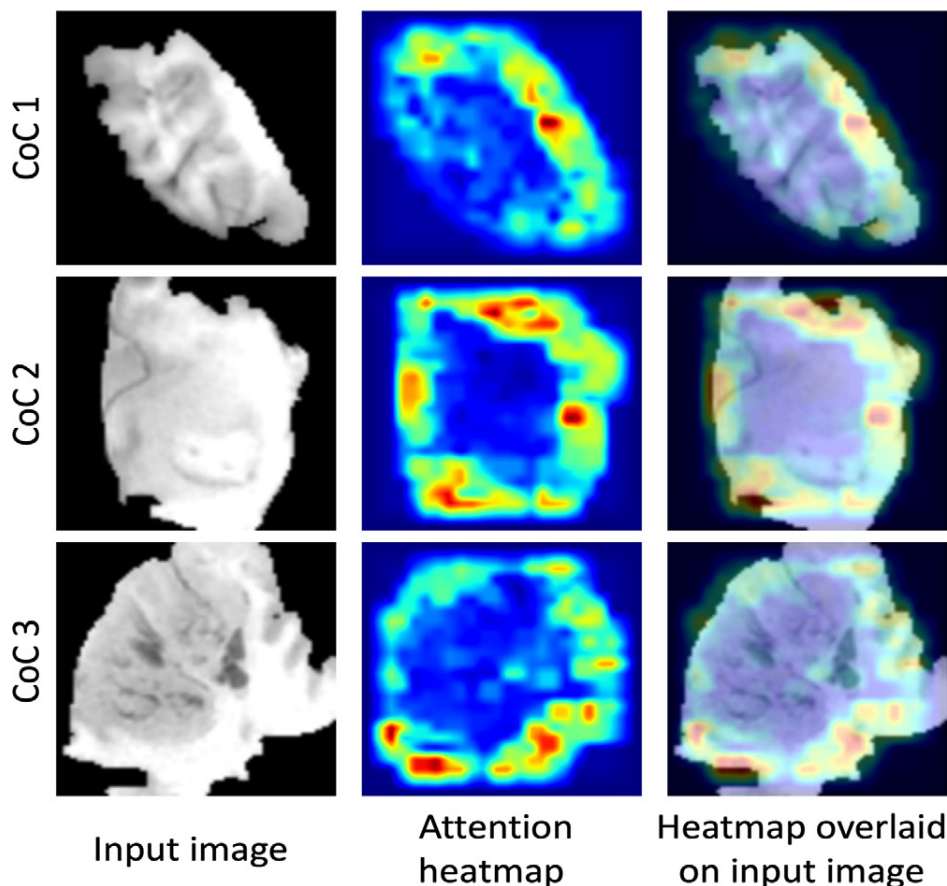
the tumor was analyzed, the imaging offers a global view that is free of this limitation. It is noted that the intratumor heterogeneity is likely a part of the reason imaging cannot predict tumor genomics with a 100% accuracy. Because the reference standard may depend on the tissue sampling strategy, it is unlikely that any predictive model can achieve perfect prediction. This limitation, caused by intratumor genomic heterogeneity, affects all studies using genomics data.

Our findings may translate to the prognosis of outcomes for patients with LGG. Specifically, we found that imaging can predict, with moderate performance, whether the tumor belongs to the CoC2 cluster or to one of the remaining genomic subtypes. The CoC2 cluster has also been shown to be highly associated with dramatically poorer survival. For example, the hazard ratio between groups CoC2 and CoC3 was 9.2 (95% CI: 4.2, 20.0), while the risk in groups CoC1 and CoC3 is similar (hazard ratio = 1.7) (1).

This shows the potential utility of the imaging-based tools to predict patient outcomes and guide treatment decisions. In addition, the task of classifying CoC2 cluster performed better than for other clusters. This could be attributed to the aggressiveness of the CoC2 cluster which is revealed in the imaging features that can be captured by a CNN model (eg, angular standard deviation of tumor shape) (6).

An interesting finding of our study was that the deep neural network that performed best was the one that utilized images of GBMs in the pretraining stage which was followed by additional training specific to LGGs. This finding illustrates that given a small set of cases such as the one used in this study, it is beneficial to allow the network to acquire general concepts of head MRIs and brain tumors even if there are some differences in the specifics of the task. It might be possible to achieve even better performance if more LGG data are available. Other recent studies have explored prediction of different relevant genomic subtypes for LGG using various methods and datasets (25–27).

Our study had some limitations, which included the limited size of the dataset as well as the fact that it was retrospectively and observationally collected. This is a common limitation in studies using comprehensive genomic analysis. While the dataset was small, it was encouraging that we were able to find meaningful relationships between genomic and imaging data.



**Figure 4:** Attention heatmaps from the network pretrained on glioblastoma dataset that indicate the parts of an image responsible for prediction. CoC = cluster of clusters.

Furthermore, to extract patches for prediction, we still needed manual segmentation masks of the tumor on each slice. Therefore, the system was not fully automatic. However, with recent advances in deep learning segmentation techniques, automatic segmentation is capable of achieving performance of an expert human reader. This implies that this step can be automated in the future, making the entire process presented in this article fully automatic.

To conclude, we were able to demonstrate that deep learning algorithms, especially those that utilize transfer learning, are able to find the association between imaging and genomics of LGGs. While the developed tool cannot yet serve as a direct replacement for genomic testing, it shows promise in aiding clinical decisions and science of lower grade gliomas.

**Author contributions:** Guarantor of integrity of entire study, M.A.M.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, M.B., A.S., M.A.M.; experimental studies, M.B., M.A.M.; statistical analysis, M.B., M.A.M.; and manuscript editing, all authors

**Disclosures of Conflicts of Interest:** M.B. disclosed no relevant relationships. E.A.A. disclosed no relevant relationships. A.S. disclosed no relevant relationships. M.A.M. Activities related to the present article: advisory role with Gradient Health. Activities not related to the present article: institution has received grant money from Bracco Diagnostics. Other relationships: disclosed no relevant relationships.

## References

1. Cancer Genome Atlas Research Network, Brat DJ, Verhaak RG, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med* 2015;372(26):2481–2498.
2. Zhang CM, Brat DJ. Genomic profiling of lower-grade gliomas uncovers cohesive disease groups: implications for diagnosis and treatment. *Chin J Cancer* 2016;35:12.
3. Eckel-Passow JE, Lachance DH, Molinaro AM, et al. Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *N Engl J Med* 2015;372(26):2499–2508.
4. Mazurowski MA. Radiogenomics: what it is and why it is important. *J Am Coll Radiol* 2015;12(8):862–866.
5. Mazurowski MA, Clark K, Czarnek NM, Shamsesfandabadi P, Peters KB, Saha A. Radiogenomic analysis of lower grade glioma: a pilot multi-institutional study shows an association between quantitative image features and tumor genomics. In: Armato SG III, Petrick NA, eds. *Proceedings of SPIE: medical imaging 2017—computer-aided diagnosis*. Vol 10134. Bellingham, Wash: International Society for Optics and Photonics, 2017; 101341T.
6. Mazurowski MA, Clark K, Czarnek NM, Shamsesfandabadi P, Peters KB, Saha A. Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with The Cancer Genome Atlas data. *J Neurooncol* 2017;133(1):27–35.
7. Yu J, Shi Z, Lian Y, et al. Noninvasive IDH1 mutation estimation based on a quantitative radiomics approach for grade II glioma. *Eur Radiol* 2017;27(8):3509–3522.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436–444.
9. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging* 2019;49(4):939–954.
10. Pedano N, Flanders AE, Scarpace L, et al. Radiology data from the cancer genome atlas low grade glioma [TCGA-LGG] collection. *Cancer Imaging Archive*. <https://wiki.cancerimagingarchive.net/display/Public/TCGA-LGG>. Published 2016. Accessed October 9, 2018.
11. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. *ArXiv* 1611.03530. [preprint] <https://arxiv.org/abs/1611.03530>. Posted November 10, 2016. Accessed DATE.
12. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw* 2018;106:249–259.
13. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. <https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks>. Published 2012. Accessed October 9, 2018.
14. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *ArXiv* 1409.1556. [preprint] <https://arxiv.org/abs/1409.1556>. Posted September 4, 2014. Accessed October 9, 2018.
15. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv* 1502.03167. [preprint] <https://arxiv.org/abs/1502.03167>. Posted February 22, 2015. Accessed October 9, 2018.
16. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011; 315–323. <http://proceedings.mlr.press/v15/glorot11a.html>.
17. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;15(1):1929–1958. <http://jmlr.org/papers/v15/srivastava14a.html>.
18. Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: an astounding baseline for recognition. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014; 806–813.
19. Azizpour H, Razavian AS, Sullivan J, Maki A, Carlsson S. From generic to specific deep representations for visual recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015; 36–45.
20. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 2014; 818–833.
21. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015; 1–9.
22. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis* 2015;115(3):211–252.
23. AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Med Phys* 2018;45(3):1150–1158.
24. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 1997;30(7):1145–1159.
25. Jakola AS, Zhang YH, Skjalsvik AJ, et al. Quantitative texture analysis in the prediction of IDH status in low-grade gliomas. *Clin Neurol Neurosurg* 2018;164:114–120.
26. Park YW, Han K, Ahn SS, et al. Prediction of IDH1-mutation and 1p/19q-codeletion status using preoperative MR imaging phenotypes in lower grade gliomas. *AJNR Am J Neuroradiol* 2018;39(1):37–42.
27. Buda M, Saha A, Mazurowski MA. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Comput Biol Med* 2019;109:218–225.

## **A.5 Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images**

## ● Original Contribution

# DEEP LEARNING-BASED SEGMENTATION OF NODULES IN THYROID ULTRASOUND: IMPROVING PERFORMANCE BY UTILIZING MARKERS PRESENT IN THE IMAGES

MATEUSZ BUDA,<sup>\*</sup> BENJAMIN WILDMAN-TOBRINER,<sup>\*</sup> KERRY CASTOR,<sup>†</sup>  
JENNY K. HOANG,<sup>\*</sup> and MACIEJ A. MAZUROWSKI<sup>\*,†</sup>

<sup>\*</sup> Department of Radiology, Duke University School of Medicine, Durham, North Carolina, USA; and <sup>†</sup> Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina, USA

(Received 17 April 2019; revised 29 September 2019; in final form 3 October 2019)

**Abstract**—Computer-aided segmentation of thyroid nodules in ultrasound imaging could assist in their accurate characterization. In this study, using data for 1278 nodules, we proposed and evaluated two methods for deep learning-based segmentation of thyroid nodules that utilize calipers present in the images. The first method used approximate nodule masks generated based on the calipers. The second method combined manual annotations with automatic guidance by the calipers. When only approximate nodule masks were used for training, the achieved Dice similarity coefficient (DSC) was 85.1%. The performance of a network trained using manual annotations was DSC = 90.4%. When the guidance by the calipers was added, the performance increased to DSC = 93.1%. An increase in the number of cases used for training resulted in increased performance for all methods. The proposed method utilizing the guidance by calipers matched the performance of the network that did not use it with a reduced number of manually annotated training cases. (E-mail: [mateusz.buda@duke.edu](mailto:mateusz.buda@duke.edu)) © 2019 World Federation for Ultrasound in Medicine & Biology. All rights reserved.

**Key Words:** Ultrasound, Deep learning, Segmentation, Thyroid nodules.

## INTRODUCTION

Thyroid nodules are extremely common and are best evaluated with ultrasound to determine whether the nodule should receive biopsy (Smith-Bindman et al. 2013; Hoang et al. 2015). Multiple groups have proposed biopsy guidelines based on the imaging appearance of nodules, but no system can achieve high specificity (Grani et al. 2019). Recently, advances in deep learning have led to algorithmic characterization of thyroid nodules, with fast and potentially accurate diagnosis (Chi et al. 2017; Ma et al. 2017) that may assist in risk stratification. An important component of these deep learning systems is the delineation of a lesion's boundaries in order for algorithms to analyze features that belong precisely to it. Segmentation of the nodule of interest is not the end goal by itself but it may be utilized in automatic assessment of ultrasound features in the Thyroid Imaging Reporting and Data System (Tessler et al. 2017),

(e.g., shape, margin, etc.). It could also facilitate other machine learning algorithms that aim to discriminate malignant and benign nodules (Buda et al. 2019). Finally, it could be used for measurement of nodule characteristics such as size.

Deep learning has been reported to be effective for segmenting a variety of medical images (De Fauw et al. 2018; Haberl et al. 2018; Wachinger et al. 2018; Zhang et al. 2019). However, generating training data for algorithms typically requires a human to manually or semi-manually outline the object of interest (i.e., provide pixel-based segmentation masks). This process is time consuming and, because it often requires the expertise of a radiologist, is also costly. Thus, large numbers of manually annotated images are not often available in medical imaging.

Ultrasound for thyroid nodules is no exception: development of deep learning algorithms requires precise segmentation that is time consuming and costly. However, thyroid ultrasound images contain additional information that could be used by deep learning algorithms to automate the segmentation process. Specifically, ultrasonographers use calipers to measure all thyroid nodules as part

Address correspondence to: Mateusz Buda, Department of Radiology, Duke University, 2424 Erwin Road, Suite 302, Durham, NC 27705, USA. E-mail: [mateusz.buda@duke.edu](mailto:mateusz.buda@duke.edu)



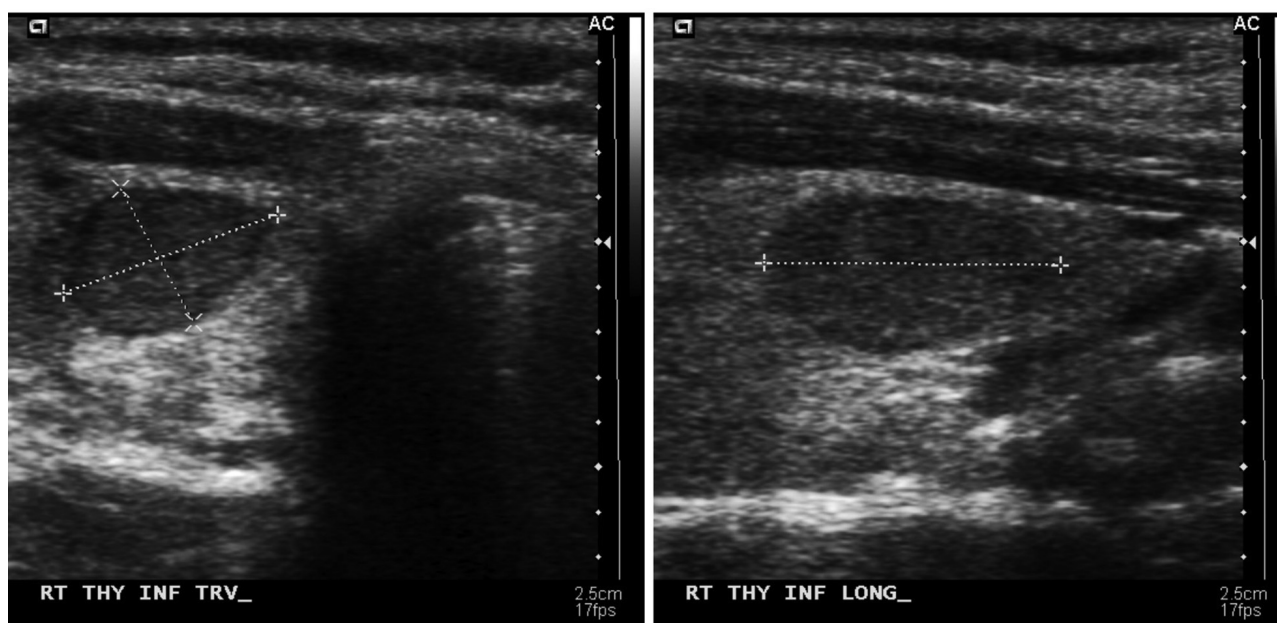


Fig. 1. Example transverse and longitudinal views of a nodule with caliper marks. Images were cropped and enlarged for better visibility.

of routine workflow. Calipers are markers (*e.g.*, small crosses or plus signs) that mark the extremes of a nodule for measurement (Fig. 1). These markers are embedded in selected images and remain on the images after the examination is completed.

With these factors in mind, our goal was to develop two deep learning methods that take advantage of calipers in thyroid ultrasound images to (i) generate approximate training data and therefore eliminate the need for manual annotations, and (ii) improve the performance of a deep learning segmentation algorithm when manual annotations are available for training. We also aimed to evaluate how changes in training set size affect algorithm performance.

## METHODS

### Data set

This institutional review board-approved, Health Insurance Portability and Accountability Act of 1996-compliant study was a retrospective analysis of ultrasound images for 1278 thyroid nodules obtained between 2006 and 2010 from 1139 patients. A waiver of consent was obtained given the retrospective nature of the study. Each nodule had exactly one image in the longitudinal plane and one in the transverse plane, resulting in 2556 total images. Images were obtained with a variety of commercially available scanners (Antares and Elegra, Siemens Healthineers, Erlangen, Germany; ATL HDI 5000 and iU22, Philips, Best, Netherlands; and Logic E9, General Electric Healthcare, Chicago, IL, USA). The data set was

randomly split into a training set of 1078 nodules (2156 images), a validation set of 107 nodules (214 images), and a test set of 200 nodules (400 images).

Each image contained either two or four calipers, markers placed by an ultrasonographer when acquiring nodule measurements as part of routine clinical practice (Fig. 1). For each image, we obtained a manually drawn nodule outline (segmentation) which was approved by a radiologist (B.W.-T., 2 y of experience in thyroid imaging). These outlines were used as the gold standard for evaluation on the test set.

### Pre-processing

In the pre-processing step, regions of interest (ROIs) containing nodules were extracted, and metadata text placed on images (information about scanner, study, patient, *etc.*) were discarded. First, we applied contrast stretching, with lower and upper pixel value limits being the first and the 99th percentile values in the original image, respectively. Next, we applied thresholding to the image at 2% of the maximum pixel intensity and extracted the bounding box of the largest connected component in the binarized image. This defined the ROI extracted from both the original image and its corresponding ground truth segmentation mask. Cropped images were zero-padded to be squares and then resized to  $320 \times 320$  pixels. The parameters for pre-processing steps were selected based on the validation set. Finally, we performed z-score normalization based on the mean and standard deviation computed on the training set.

### BASE SEGMENTATION MODEL

The network architecture of the segmentation model was a fully convolutional encoder–decoder U-Net (Ronneberger *et al.* 2015). It was built based on blocks comprising convolutional filters, rectified linear unit activation function, batch normalization and max-pooling or deconvolutional layers in the encoder and decoder parts, respectively (Fig. 2). The numbers of filters in convolutional and corresponding deconvolutional layers were 32, 64, 128 and 256. Skip-connections with concatenation operation were used between encoder and decoder blocks. Input images and output masks were of size  $320 \times 320$  pixels. During the training, the optimized function was Dice similarity loss (Sudre *et al.* 2017). The maximum number of training epochs was set to 100. To prevent overfitting, an early stopping strategy was

used with a patience of 10 epochs (Caruana *et al.* 2001), and model selection was based on the best loss computed on the validation set. In addition, during training, we applied data augmentation to the training examples in the form of random rotation by  $90^\circ$ , horizontal and vertical shift by 32 pixels, as well as scale by 10%.

#### Segmentation approaches

##### Approach 1: Using approximate masks for training.

This technique used calipers within each image to automatically generate an approximate segmentation mask. Manual nodule segmentations were not used for training. To obtain approximate segmentation masks, calipers were automatically identified using a detection model Faster R-CNN (Ren *et al.* 2015) with ResNet-101 backbone (He *et al.* 2016). It was trained on 7858 annotated calipers

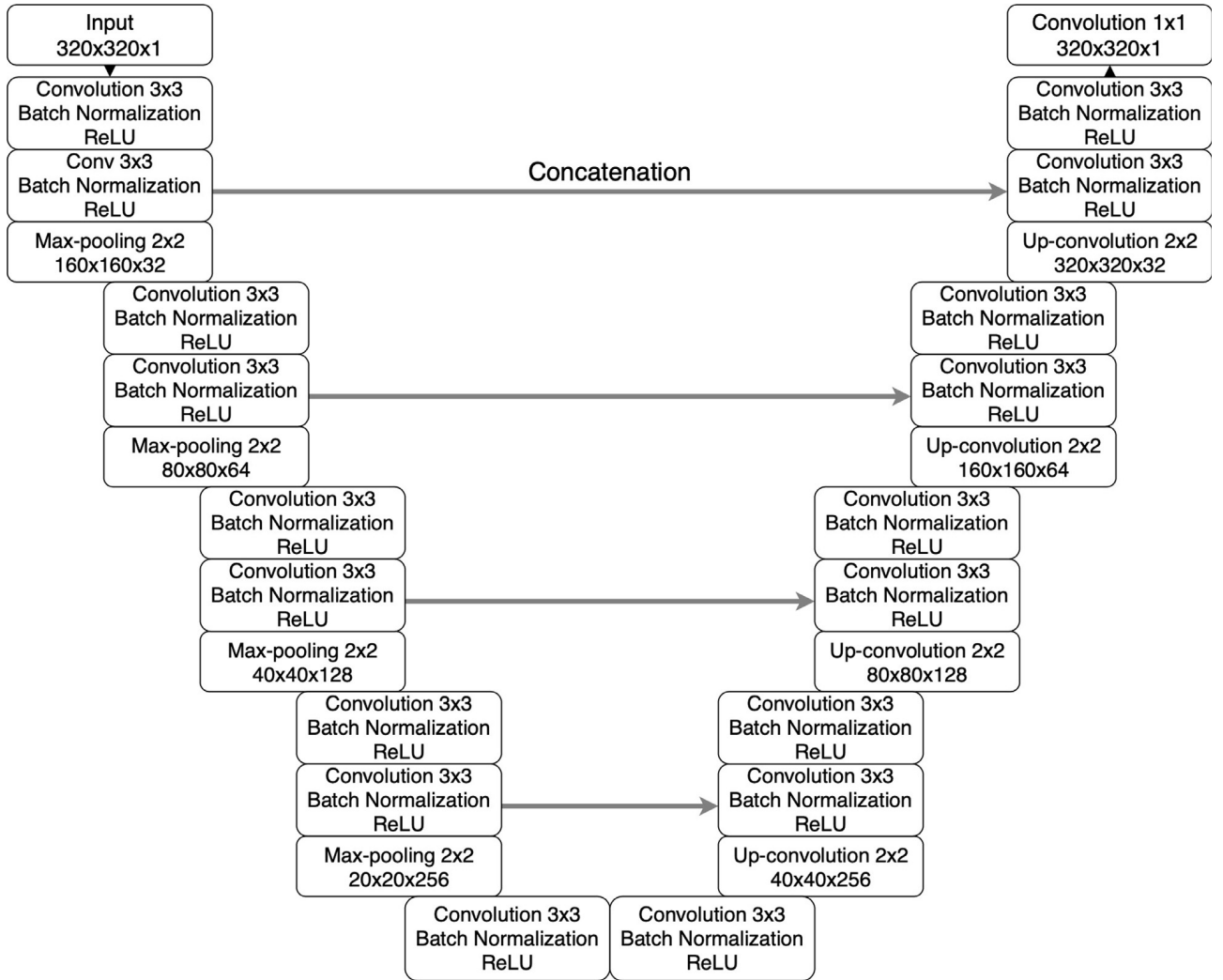


Fig. 2. Base U-Net architecture used in experiments. It comprised four encoding and four decoding blocks. Each block contained convolutional layer, batch normalization and ReLU activation, repeated two times and followed by max-pooling and up-convolutional layer in the encoder and decoder parts, respectively.

in all 2556 images in our data set. Using 10-fold cross-validation, we generated predicted locations of calipers for all images. The precision of this caliper detection method, evaluated with 10-fold cross-validation, was 99.77% at 0.5 intersection over union threshold (Lin et al. 2014). Then, these predicted caliper locations were used to produce approximate nodule segmentations by applying cubic interpolation connecting calipers. When there were only two calipers, we added extra points assuming a circular nodule or, if a circle did not fit within the image, an elliptical nodule shape was used (Fig. 3). Once the approximate masks are generated, they are used to train the basic segmentation model based on a fully convolutional encoder–decoder model as described above.

*Approach 2: Using precise masks for training.* This technique used manual segmentation masks for training instead of approximate ones. The process of obtaining these precise masks required a radiologist with practice in reporting thyroid nodules, and it took approximately 10 s for one image. The model architecture, training procedure and training hyperparameters were the same as for approach 1.

*Approach 3: Using a combination of approximate and precise masks for training.* This approach used both manual and approximate masks. However, precise masks were used as ground truth for training, and approximate masks were used only as an additional input channel together with original nodule ultrasound images.

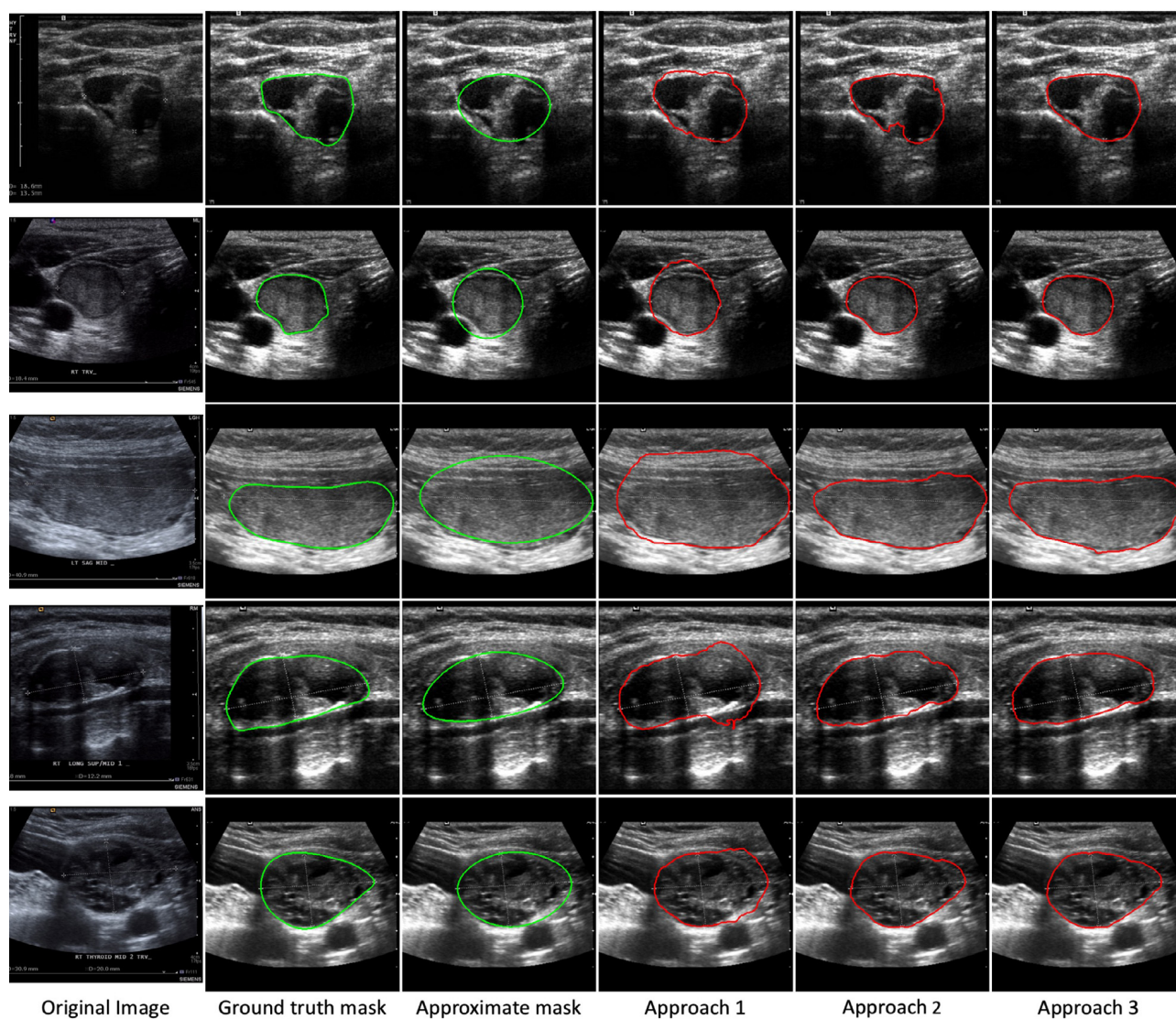


Fig. 3. Qualitative results comparing segmentations generated with deep learning model and ground truth masks for example images from the test set.



Approximate masks were generated fully automatically as described for approach 1. The training procedure was the same as for previous approaches.

### Evaluation

All three approaches were evaluated on the test set of 200 cases (400 images). The results of the deep learning-based segmentations were compared with manually outlined masks. The Dice similarity coefficient (DSC) averaged across all images was used as the performance metric. Statistical tests for comparison of results were performed using bootstrapping.

Additionally, each experiment was run on 10 subsets of the training set with varying numbers of training examples to investigate the effect of training set size on segmentation performance. The smallest training set size comprised 10% of the original training set and was increased by 10% in each subset until it reached 100% of the training set. For more accurate performance estimation, we repeated training of each network 10 times starting with randomly initialized weights. This experimental setting resulted in 300 models in total, all trained from scratch.

## RESULTS

### Effects of type of ground truth on segmentation performance

When training occurred on the entire training set, use of approximate masks for development (approach 1) resulted in a good performance, with a median DSC of 85.1% on the test set. When precise manual segmentations were used as the ground truth instead of approximate masks (approach 2), the trained model achieved a higher median DSC of 90.4% ( $p < 0.001$ ). In approach 3, the trained network was guided by the approximate mask for the image of interest (used as input). In this setting, performance increased to a median DSC of 93.1%. This is an improvement of 8% over approach 1 ( $p < 0.001$ ) and 2.7% over approach 2 ( $p < 0.001$ ).

### Effects of training set size on segmentation performance

Performance increased as training set size increased until about 50%–70% of the original training set was used. The model training with additional input channel (approach 3), even when training using only 20% of the cases from the original training set, outperformed all other models trained on the entire training set. Moreover, this setting resulted in the most stable training, with the difference in DSC for all trained models and for all subsets of the training set within 1%. Figure 4 contains boxplots with results from all experiments.

Regarding the comparison between approximate and precise masks, use of 10% of precise masks for training resulted in a DSC similar to that obtained

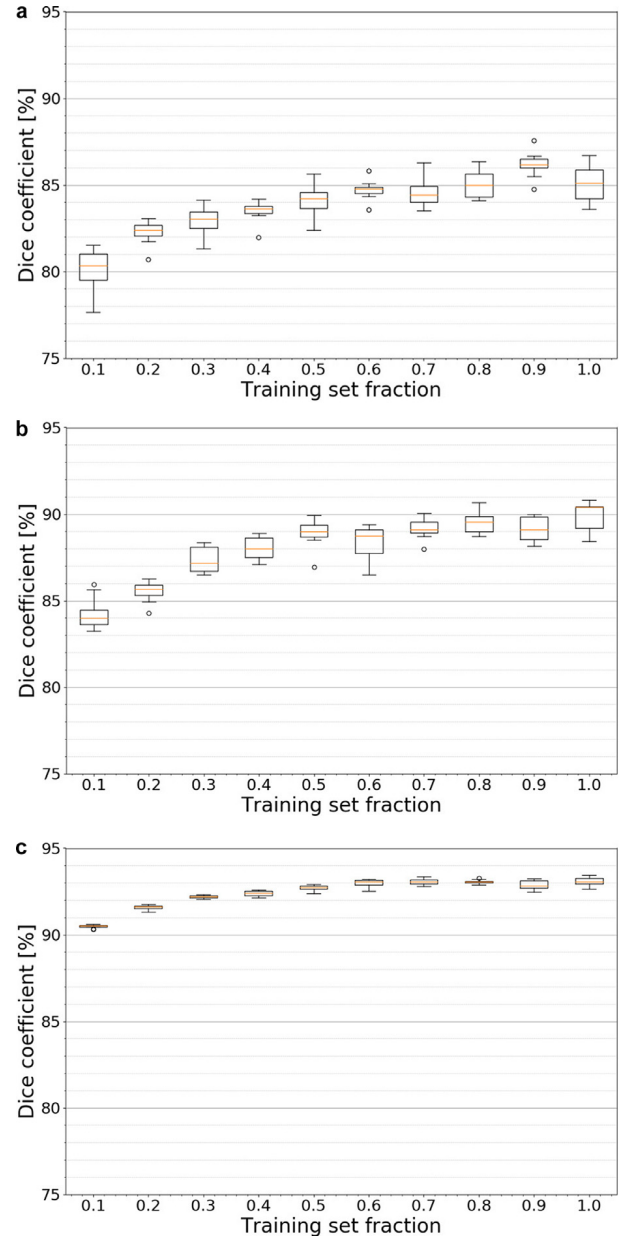


Fig. 4. Boxplots revealing mean DSC achieved by a model trained using (a) automatically generated approximate GT masks (approach 1), (b) precise GT masks (approach 2) and (c) precise GT masks and automatically generated approximate masks as additional input channel. DSC = Dice similarity coefficient; GT = ground truth.

using 50% of approximate masks (84% DSC), whereas use of 20% of precise masks resulted in a DSC similar to that obtained using 100% of the training set of approximate masks (85% DSC). It means that in these settings, approximately five times more automatic approximate masks were needed than precise ones to achieve similar results. However, we estimate the time needed to obtain one precise mask was about 10 s for a trained radiologist, compared with automatic masks

that were generated in about 1 s and without any radiologist supervision.

## DISCUSSION

Segmenting lesions in medical imaging for deep learning purposes is a time-consuming process, and radiologists or trained professionals may spend hours outlining lesions so that deep learning tools may be developed. Our study proposed and evaluated deep learning approaches to efficiently and accurately segment thyroid nodules on ultrasound. We found that exceptionally high performance can be achieved when applying these approaches. Specifically, we utilized calipers inherent in ultrasound images both to improve performance and to create an automated segmentation system that could reduce the manual labor needed to create a data set for development of deep learning models. The proposed approaches are also model agnostic and can facilitate other existing segmentation methods.

Our first approach to automated thyroid segmentation directly addresses this issue, as it allows for training of a fairly accurate (DSC = 85.1%) segmentation algorithm without using any manual segmentation data. Similar performance for approaches relying on manual annotations was reached when using 10% or 20% of total training examples. Please note that 20% of all training data corresponds to approximately 216 nodules, which would require annotation of 432 individual images (2 images per nodule). This translates to a substantial time commitment that could be avoided using our method, which uses automatic approximate annotations.

In the approach using only approximate masks, once the caliper detection algorithm was developed, the process of generating masks for training was completely automatic. This is of high significance because no manual segmentation is required for training, and therefore, the network can be developed without radiologist-provided input. If such an algorithm were not available, manual marking of caliper position could still be performed, which maintains the benefits of demanding less time than creating an enclosed outline for each nodule and of not requiring a trained radiologist.

Although calipers proved beneficial for development of a fully automated segmentation algorithm, they were also helpful when manual segmentation data were used. Specifically, the calipers were detected to provide a “first guess” mask as an additional network input. This approach provided a notable increase in performance from 90.4%–93.1% ( $p < 0.001$ ). Alternatively, when only 10% of the annotated cases were available, the proposed method achieved a performance of 90.5%. This means that the proposed method is capable of achieving the same performance as the best standard network trained on manually annotated data at a dramatically

reduced cost in terms of annotations. An interesting finding is that the proposed caliper-based guidance has resulted in a dramatic decrease in the variability of performance of different trained networks. Such stability is a very desired effect in practical settings.

Our study had limitations. First, some thyroid nodules have ill-defined margins, which makes them difficult (if not impossible) to accurately outline, even by an expert reader. In addition, thyroid nodules that are very large can also have margins difficult to annotate because the entire nodule could not fit within a single image. These properties, in turn, might result in high inter-observer variability. Because our segmentation masks were provided by a single reader, there is a risk of biased segmentation results and overfitting to a specific reader. On the other hand, our data set was not curated to include only nodules with well-defined margins. This might limit performance, but it increases generalizability.

Because thyroid ultrasound is a very common imaging exam, an abundance of data is available in hospital picture archiving and communication systems. As one of the methods proposed in this article does not require any additional annotation other than what is already present in the images, a future study could evaluate the performance of a model trained using the approximate, caliper-based masks with a 10- to 100-fold larger data set. Although we observed some saturation of the performance with the increasing number of cases, a notable increase in performance is possible with a much larger training sample. There is some potential that with a very large sample, the performance of the networks trained on manually generated and approximate, caliper-based masks converge, although this hypothesis requires additional validation.

In the past, other computer vision algorithms were used for segmentation of thyroid nodules in ultrasound images. In [Iakovidis et al. \(2007\)](#), the authors proposed a level set approach based on a variable background active contour model ([Maroulis et al. 2007](#)) with hyper-parameters tuned using genetic algorithms. As evaluated on 45 images, average overlap with ground truth segmentations was 92.5%; however, this method was applicable only to hypo-echoic nodules. This limitation was overcome in the joint echogenicity–texture active contour model ([Savelonas et al. 2008](#)) which achieved mean DSCs of 96.3% and 95.5% on 38 hypo-echoic and 36 iso-echoic nodules, respectively. Finally, a level set approach based on neutrosophic  $L$ -means clustering, proposed in [Koundal et al. \(2016\)](#), obtained a mean DSC of 94.2% on a data set of 42 images containing hypo-echoic and hyper-echoic nodules. These segmentation methods were evaluated on relatively small data sets (compared with the 400 test images in our study), and some of them apply only to nodules of a specific echogenicity. In addition, evaluation was performed based on the same set of cases

that were used for algorithm development, which could significantly bias the results.

## CONCLUSIONS

In this study, we found that caliper marks placed on ultrasound images can be employed in development of a deep learning-based segmentation system in two ways. First, the technique was used to automatically generate a data set of approximate masks for training. Second, these generated approximate masks, used as additional input to the network, notably improved segmentation performance.

*Conflict of interest disclosure*—Maciej Mazurowski declares an advisory relationship with Gradient Health. All other authors declare that they have no conflicts of interest.

## REFERENCES

- Buda M, Wildman-Tobriner B, Hoang JK, Thayer D, Tessler FN, Middleton WD, Mazurowski MA. Management of thyroid nodules seen on US Images: Deep learning may match performance of radiologists. *Radiology* 2019;292(3).
- Caruana R, Lawrence S, Giles L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in Neural Information Processing Systems 13 - Proceedings of the 2000 Conference (NIPS 2000)*. In: *Neural information processing systems Foundation*; 2001.
- Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *J Digit Imaging* 2017;30:477–486.
- De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, Glorot X, O'Donoghue B, Visentin D. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342.
- Grani G, Lamartina L, Ascoli V, Bosco D, Biffoni M, Giacomelli L, Maranghi M, Falcone R, Ramundo V, Cantisani V. Reducing the number of unnecessary thyroid biopsies while improving diagnostic accuracy: Toward the “right” TIRADS. *J Clin Endocrinol Metab* 2019;104:95–102.
- Haberl MG, Churas C, Tindall L, Boassa D, Phan S, Bushong EA, Madany M, Akay R, Deerinck TJ, Peltier ST. CDeep3 M—Plug-and-Play cloud-based deep learning for image segmentation. *Nat Methods* 2018;15:677.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Conf Comput Vis Pattern Recognit* 2016;770–778.
- Hoang JK, Langer JE, Middleton WD, Wu CC, Hammers LW, Cronan JJ, Tessler FN, Grant EG, Berland LL. Managing incidental thyroid nodules detected on imaging: White paper of the ACR Incidental Thyroid Findings Committee. *J Am Coll Radiol* 2015;12:143–150.
- Iakovidis DK, Savelonas MA, Karkanis SA, Maroulis DE. A genetically optimized level set approach to segmentation of thyroid ultrasound images. *Appl Intell* 2007;27:193–203.
- Koundal D, Gupta S, Singh S. Automated delineation of thyroid nodules in ultrasound images using spatial neutrosophic clustering and level set. *Appl Soft Comput* 2016;40:86–97.
- Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. *Eur Conf Comput Vis* 2014;740–755.
- Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. *Ultrasonics* 2017;73:221–230.
- Maroulis DE, Savelonas MA, Iakovidis DK, Karkanis SA, Dimitropoulos N. Variable background active contour model for computer-aided delineation of nodules in thyroid ultrasound images. *IEEE Trans Inf Technol Biomed* 2007;11:537–543.
- Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst* 2015;91–99.
- Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *Int Conf Med Image Comput Comput Interv* 2015;234–241.
- Savelonas MA, Iakovidis DK, Legakis I, Maroulis D. Active contours guided by echogenicity and texture for delineation of thyroid nodules in ultrasound images. *IEEE Trans Inf Technol Biomed* 2008;13:519–527.
- Smith-Bindman R, Lebda P, Feldstein VA, Sellami D, Goldstein RB, Brasic N, Jin C, Kornak J. Risk of thyroid cancer based on thyroid ultrasound imaging characteristics: Results of a population-based study. *JAMA* 2013;173:1788–1795.
- Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, held in Conjunction with MICCAI 2017*. Québec City, QC, Canada, September 14. Springer; 2017. p. 240–248.
- Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC, Hammers LW, Hamper UM, Langer JE, Reading CC, Scott LM, Stavros AT. ACR Thyroid Imaging, Reporting and Data system (TI-RADS): White paper of the ACR TI-RADS committee. *J Am Coll Radiol* 2017;14:587–595.
- Wachinger C, Reuter M, Klein T. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *Neuroimage* 2018;170:434–445.
- Zhang J, Saha A, Zhu Z, Mazurowski MA. Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics. *IEEE Trans Med Imaging* 2019;38:435–447.

## **A.6 Management of thyroid nodules seen on US images: deep learning may match performance of radiologists**



# Management of Thyroid Nodules Seen on US Images: Deep Learning May Match Performance of Radiologists

Mateusz Buda, MSc • Benjamin Wildman-Tobriner, MD • Jenny K. Hoang, MBBS, MHS • David Thayer, PhD, MD • Franklin N. Tessler, MD • William D. Middleton, MD • Maciej A. Mazurowski, PhD

From the Department of Radiology, Duke University School of Medicine, 2424 Erwin Road, Suite 302, Durham, NC 27705 (M.B., B.W.T., J.K.H., M.A.M.); Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, Mo (D.T., W.D.M.); Department of Radiology, University of Alabama at Birmingham, Birmingham, Ala (F.N.T.); and Department of Electrical and Computer Engineering, Duke University, Durham, NC (M.A.M.). Received June 5, 2018; revision requested July 26; revision received April 23, 2019; accepted May 29. Address correspondence to M.B. (e-mail: [mateusz.buda@duke.edu](mailto:mateusz.buda@duke.edu)).

M.B., B.W.T., J.K.H., and M.A.M. supported by the Putman Innovation Award.

Conflicts of interest are listed at the end of this article.

Radiology 2019; 292:695–701 • <https://doi.org/10.1148/radiol.2019181343> • Content codes:  

**Background:** Management of thyroid nodules may be inconsistent between different observers and time consuming for radiologists. An artificial intelligence system that uses deep learning may improve radiology workflow for management of thyroid nodules.

**Purpose:** To develop a deep learning algorithm that uses thyroid US images to decide whether a thyroid nodule should undergo a biopsy and to compare the performance of the algorithm with the performance of radiologists who adhere to American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS).

**Materials and Methods:** In this retrospective analysis, studies in patients referred for US with subsequent fine-needle aspiration or with surgical histologic analysis used as the standard were evaluated. The study period was from August 2006 to May 2010. A multitask deep convolutional neural network was trained to provide biopsy recommendations for thyroid nodules on the basis of two orthogonal US images as the input. In the training phase, the deep learning algorithm was first evaluated by using 10-fold cross-validation. Internal validation was then performed on an independent set of 99 consecutive nodules. The sensitivity and specificity of the algorithm were compared with a consensus of three ACR TI-RADS committee experts and nine other radiologists, all of whom interpreted thyroid US images in clinical practice.

**Results:** Included were 1377 thyroid nodules in 1230 patients with complete imaging data and conclusive cytologic or histologic diagnoses. For the 99 test nodules, the proposed deep learning algorithm achieved 13 of 15 (87%; 95% confidence interval [CI]: 67%, 100%) sensitivity, the same as expert consensus ( $P > .99$ ) and higher than five of nine radiologists. The specificity of the deep learning algorithm was 44 of 84 (52%; 95% CI: 42%, 62%), which was similar to expert consensus (43 of 84; 51%; 95% CI: 41%, 62%;  $P = .91$ ) and higher than seven of nine other radiologists. The mean sensitivity and specificity for the nine radiologists was 83% (95% CI: 64%, 98%) and 48% (95% CI: 37%, 59%), respectively.

**Conclusion:** Sensitivity and specificity of a deep learning algorithm for thyroid nodule biopsy recommendations was similar to that of expert radiologists who used American College of Radiology Thyroid Imaging and Reporting Data System guidelines.

© RSNA, 2019

Online supplemental material is available for this article.

Imaging with US remains an accurate method to guide recommendation for management of thyroid nodules (1), although interpretation variability and overdiagnosis represent continual challenges (2,3). To help radiologists improve consistency, several organizations have developed imaging criteria to aid in the selection of nodules recommended for fine-needle aspiration (FNA) biopsy. In 2017, the American College of Radiology (ACR) published its Thyroid Imaging Reporting and Data System (TI-RADS) (4). Similar to its predecessors, ACR TI-RADS is on the basis of US features and maximum nodule size. ACR TI-RADS has been shown to increase accuracy and specificity compared with other systems (5), enhance report quality, and improve recommendations for management (6).

Despite these potential benefits, certain barriers may prevent radiologists from adopting or using ACR TI-RADS. First, a high interobserver variability among radiologists'

interpretations has been shown with the system ( $\kappa = 0.51$ ) (2). Such variability may lead to inconsistent recommendations for nodule management between readers. Second, evaluating multiple nodules (with multiple features per nodule) can be labor intensive and could be more time consuming for some radiologists. Any practice that adds time to an already busy radiology workflow could serve as a disincentive for adopting best practices.

Because of these types of challenges, the medical community has started to use deep learning (7). Deep learning represents an approach to artificial intelligence that has been increasingly applied throughout medicine, with emerging applications in fields such as dermatology (8), ophthalmology (9), and radiology (10,11). Recent deep learning research in radiology has shown algorithm performance comparable to radiologists (12), and as the field continues to grow the variety and number of possible uses for deep



## Abbreviations

ACR = American College of Radiology, AUC = area under the receiver operating characteristic curve, CI = confidence interval, FNA = fine-needle aspiration, TI-RADS = Thyroid Imaging Reporting and Data System

## Summary

A deep convolutional neural network that uses American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) features for training achieved similar sensitivity and specificity for recommending biopsy for thyroid nodules observed at US compared with radiologists who use ACR TI-RADS.

## Key Points

- For discriminating malignant and benign nodules, deep learning achieved an area under the receiver operating characteristic curve (AUC) of 0.87 (95% confidence interval [CI]: 0.76, 0.95), which is comparable to the AUC of 0.91 (95% CI: 0.82, 0.97) for a consensus of three American College of Radiologists (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) committee experts ( $P = .42$ ) and the mean AUC of 0.82 (95% CI: 0.73, 0.90) for nine individual radiologists ( $P = .38$ ).
- Our deep learning system achieved 52% specificity and 87% sensitivity in recommending biopsy for thyroid nodules compared with 51% specificity ( $P = .91$ ) and 87% sensitivity ( $P > .99$ ) from a consensus of three ACR TI-RADS committee experts.

learning continue to increase. Some of the challenges of thyroid US interpretation and reporting data systems such as ACR TI-RADS represent problems that may be solved through deep learning applications.

The aim of our study was to design a deep learning algorithm that uses thyroid US images to decide whether a thyroid nodule should undergo a biopsy. We also aimed to compare the performance of the algorithm to that of radiologists with varying expertise who adhere to ACR TI-RADS interpretation criteria.

## Materials and Methods

### Study Population

In this institutional review board–approved, Health Insurance Portability and Accountability Act–compliant study, we retrospectively analyzed a data set of thyroid nodules. The initial population included 1631 nodules in 1439 adult patients from a single institution who underwent diagnostic thyroid US examinations and US-guided FNA of a focal thyroid nodule between August 2006 and May 2010. It was refined by excluding 203 nodules in 172 patients who had initial nondiagnostic or indeterminate cytologic results and without subsequent cytologic or histologic diagnoses. Nodules in which images on one or both orthogonal planes were missing ( $n = 15$ ) were also excluded. In addition, to facilitate nodule detection (based on a method that uses calipers), cases that did not contain images with proper caliper measurement marks (at least one caliper measurement on one plane and two on the other) were excluded ( $n = 36$ ). This resulted in 1377 nodules from 1230 patients. In the final sets for the analysis, there were 1278 nodules from 1139 patients in the training set and 99 nodules from 91 patients in the test set (Fig 1). The 99 test nodules were not

used during algorithm development. They were analyzed by multiple readers in a previous study (5).

The US examinations were performed by using a variety of commercially available units (Antares and Elegra, Siemens Healthineers, Erlangen, Germany; ATL HDI 5000 and iU22, Philips, Best, the Netherlands; and Logic E9, GE Healthcare, Waukesha, Wis) equipped with 5–15-MHz linear array transducers.

### Pathologic Ground Truth

FNA samples were obtained during standard clinical workflow and cytologic results were reviewed by pathology faculty at the institution (Washington University, St Louis, Mo). Determination of benignity or malignancy was made by using FNA results or, when available, surgical specimens. For FNA, five categories were used: malignant, suspicious for malignancy, indeterminate, benign, and nondiagnostic. We included nodules that were malignant or benign on the basis of initial FNA results or if a nodule underwent repeated FNA or surgical resection that subsequently provided confirmation of malignancy or benignity.

### Image Annotation

All images in the training set were interpreted by one of two radiologists who were blinded to pathologic results. These two radiologists were later on the ACR TI-RADS steering committee and helped to develop ACR TI-RADS. The first reader (W.D.M.) had 22 years of experience and the second reader had 20 years of experience in thyroid imaging. By following the ACR TI-RADS lexicon, the readers assigned features for nodule composition, echogenicity, margins, and echogenic foci. For the echogenicity category, the readers classified 243 nodules as moderate to markedly hypoechoic, which was not compatible with the ACR TI-RADS lexicon. For these cases, a third reader (B.W.T., a board-eligible radiology fellow with specialty practice in thyroid imaging and 5 years of experience) reviewed the echogenicity feature and modified it by using the original assignment and additional imaging review. This reader also evaluated nodules for the shape feature. Eventually, all 1377 nodules were appropriately assigned to all five ACR TI-RADS categories.

Annotations for the five ACR TI-RADS feature categories for the test nodules were performed by 12 radiologists in December 2016, before the publication of ACR TI-RADS, with the readers blinded to the pathologic results. These interpretations were on the basis of images obtained on transverse and longitudinal planes, and video clips obtained on at least one plane displayed to the readers on standard computer monitors by using a website interface. Independent interpretations by three radiologists who were experts on the ACR TI-RADS committee, one of whom is a coauthor (F.N.T.), were combined into an expert consensus by using majority vote. These radiologists had between 26 and 34 years of posttraining experience.

Among the remaining nine readers, one reader (W.D.M.) had 22 years of experience and also interpreted the training cases. The other eight radiologists reported thyroid US in their clinical practice but had no knowledge of the management recommendations in ACR TI-RADS. This group included two academic radiologists with subspecialty training in US

and 20 and 32 years of practice experience, respectively. The six remaining radiologists from this group were from private practices with fellowship training in neuroradiology, women's imaging, and nuclear medicine, with experience ranging from 3 to 32 years.

On the basis of feature assessments for the five ACR TI-RADS categories from each reader, we first computed a total number of points per nodule and corresponding ACR TI-RADS risk levels. Then, according to ACR TI-RADS guidelines, we retrospectively decided whether a nodule would qualify for FNA and follow-up on the basis of nodule size and ACR TI-RADS risk level.

### Deep Learning Algorithm

Our proposed deep learning algorithm had three main stages: nodule detection followed by prediction of malignancy and risk-level stratification. Figure 2 shows these stages and how they are connected. A complete description of all the components of the deep learning algorithm are provided in Appendix E1 (online).

For nodule detection, we first obtained a bounding box of a nodule by enclosing calipers included in every image (used in clinical practice for nodule measurement). To detect the calipers, we trained a Faster Region-based Convolutional Neural Network detection algorithm (13). After detecting the calipers on the US image, we extracted a square image with a fixed size margin of 32 pixels enclosing the corresponding nodule, resized the image to  $160 \times 160$  pixels, and applied preprocessing (Appendix E1 [online]).

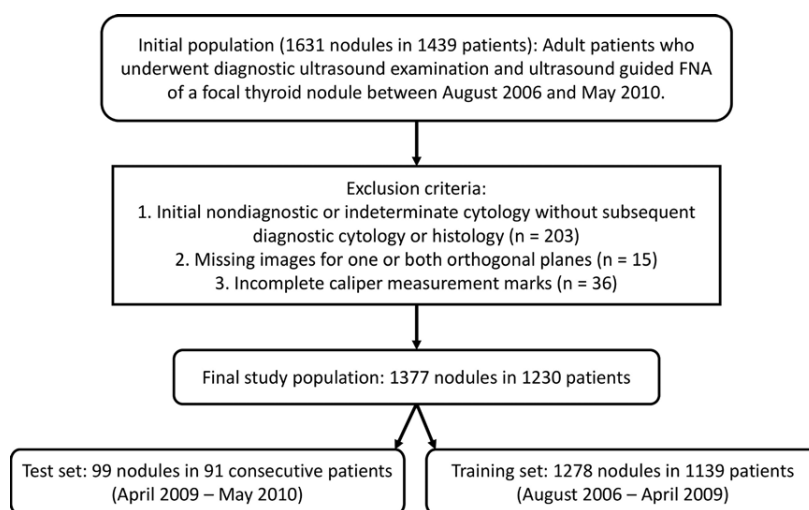
For classification, we trained a custom, multitask deep convolutional neural network. The tasks used for training were presence or absence of malignancy and all of the ACR TI-RADS features across the five categories (composition, echogenicity, shape, margin, and echogenic foci). The architecture of our common representation extraction network is shown in Figure 3. Source code of the model is available at the following link: <https://github.com/MaciejMazurowski/thyroid-us>.

During inference, we stratified the probability of malignancy returned by the network into risk levels referred to as deep learning risk levels (ie, DL2–DL5), modeled after the ones defined in ACR TI-RADS (ie, TR2–TR5).

Use of the deep learning risk level and a nodule's size resulted in a recommendation for FNA and follow-up. The size thresholds for FNA and follow-up recommendation were the same as in ACR TI-RADS. We used this step to choose the appropriate point on a receiver operating characteristic curve that considers nodule size and results in clinically relevant decisions.

### Evaluation

We evaluated our deep learning algorithm and compared it with the performance of radiologists in two steps (Fig 4). First, we compared the performance of the algorithm to human readers for discriminating benign and malignant nodules alone by using the area under the receiver operating characteristic curve (AUC). This is the first and principal step of our algorithm and the ACR TI-RADS, and it does not involve nodule size. The AUC for the deep learning algorithm was calculated by using the likelihood of malignancy returned by model, and the AUC for radiologists used the total number of points computed with ACR TI-RADS. Then, for the second step, we evaluated the performance of the entire system in terms of sensitivity and specificity for recommendation of FNA and follow-up that in addition to the first step involves size-based thresholding. This two-step evaluation allows for isolating the predictive performance that is purely on the basis of the image from the final size-based recommendation step that aims to relate to the risk that malignant nodules



**Figure 1:** Flowchart of inclusion criteria for initial population and exclusion criteria for the final study population. FNA = fine-needle aspiration.



**Figure 2:** Flowchart of the three main processing stages of our deep learning algorithm. CNN = convolutional neural network, R-CNN = Region-based CNN, ROI = region of interest.

of different sizes pose to patients.

We performed validation of the performance of the deep learning classifier in two ways: by using a 10-fold cross-validation with our training set by pooling predictions from all 10 nonoverlapping folds and by using a hold-out test set of 99 cases. For the training set, AUC of the deep learning algorithm was compared with that of a single radiologist. On the test set, we compared the deep learning with consensus of the three ACR TI-RADS committee members and the nine other radiologists. Statistical tests for all comparisons were performed with bootstrapping.

## Results

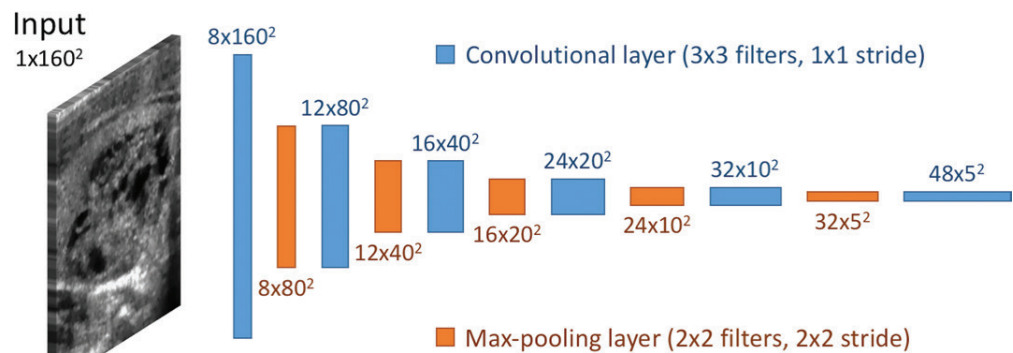
### Study Population

The total number of malignant nodules was 142 (of 1377 nodules; 10.3%); there were 127 malignant nodules (of 1278 nodules; 9.9%) in the training set and 15 malignant nodules (of 99 nodules; 15%) in the test set (Table 1). The prevalence of malignant nodules between the training and test sets was not statistically significant ( $P = .09$ ). The mean maximum nodule size for all cases was 2.6 cm (2.6 cm in the training set and 2.7 cm in the test set;  $P = .53$ ).

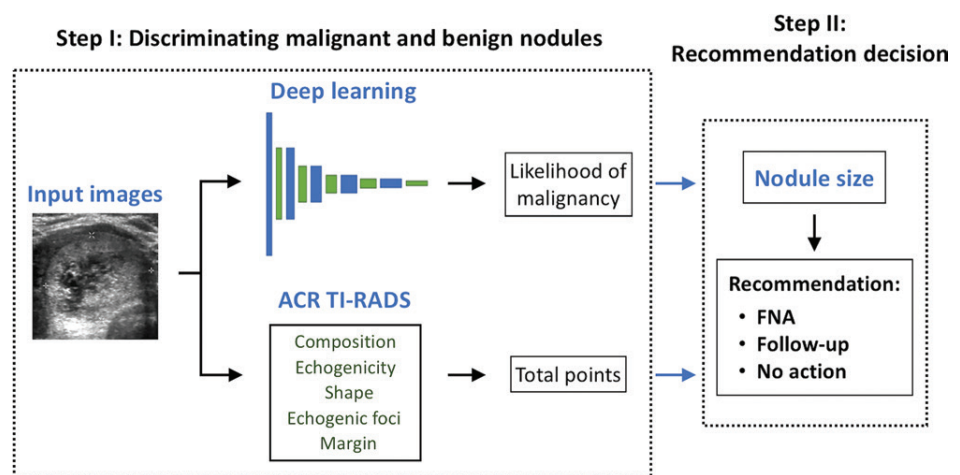
### Comparison of Deep Learning and Radiologists

For the training set of 1278 nodules, evaluated by using 10-fold cross-validation, the deep learning algorithm achieved an AUC of 0.78 (95% confidence interval [CI]: 0.74, 0.82) compared with 0.80 (95% CI: 0.76, 0.84;  $P = .44$ ) for a single ACR TI-RADS committee radiologist by using ACR TI-RADS (Fig 5a).

For the test set for discriminating malignant and benign nodules, deep learning achieved an AUC of 0.87 (95% CI: 0.76, 0.95), which is comparable ( $P = .42$ ) to that of expert consensus (0.91; 95% CI: 0.82, 0.97). The mean AUC of the nine radiologists was 0.82 (95% CI: 0.73, 0.90; not significantly lower than for deep learning,  $P = .38$ ); the lowest AUC was 0.76 (95% CI: 0.63, 0.88) and the highest AUC was 0.85 (95% CI: 0.76, 0.94). The performance of eight of the nine individual radiologists was worse than that of deep learning; however, these differences were not statistically significant ( $P > .08$ ). The score of each reader is provided in Table 2 and



**Figure 3:** Convolutional neural network architecture of the network for shared representation extraction.



**Figure 4:** A diagram of the two-step decision-making process for management of thyroid nodules. ACR = American College of Radiology, FNA = fine-needle aspiration, TI-RADS = Thyroid Imaging Reporting and Data System.

the mean receiver operating characteristic curve is shown in Figure 5b.

After applying risk level stratification and size thresholds for FNA recommendation according to ACR TI-RADS, the sensitivity of the proposed deep learning algorithm was 13 of 15 (87%; 95% CI: 67%, 100%), the same as the expert consensus sensitivity of 13 of 15 (87%; 95% CI: 67%, 100%). For the nine radiologists, sensitivity ranged from 11 of 15 (73%) to 14 of 15 (93%). The differences between sensitivity of deep learning and radiologists were not statistically significant ( $P > .43$ ). In terms of specificity, deep learning achieved 44 of 84 (52%; 95% CI: 41%, 63%), which was higher (although not significantly;  $P = .91$ ) than expert consensus (43 of 84; 51% [95% CI: 41%, 62%]) and seven of the nine radiologists with specificity ranging from 24 of 84 (29%) to 59 of 84 (70%). The differences between specificity of deep learning and two of these seven radiologists (reader 2 and reader 8) were statistically significant ( $P < .001$  and  $P = .042$ , respectively). The mean sensitivity and specificity for all nine radiologists was 83% (95% CI: 64%, 98%) and 48% (95% CI: 37%, 59%), respectively; both mean sensitivity and mean specificity were lower than for the deep learning algorithm (sensitivity and specificity,  $P = .68$  and  $.45$ , respectively). Sensitivity and specificity



for FNA recommendation by all readers is provided in Table 2. Of the nodules that were misclassified by deep learning (42%; 95% CI: 33%, 53%), the nine radiologists misclassified an average of 72% (95% CI: 59%, 83%) of nodules. However, of the nodules misclassified by radiologists (average error rate, 47%; 95% CI: 37%, 56%), deep learning misclassified 66% (95% CI: 53%, 77%) of nodules. This shows a notable overlap in the misclassified cases and somewhat lower misclassification rate by the deep learning algorithm compared with that of the radiologists.

When recommending follow-up for nodules stratified into risk levels and when using size thresholds according to ACR TI-RADS, deep learning performed similarly to the radiologists. Its sensitivity was 14 of 15 (93%; 95% CI: 78%, 100%). Expert consensus did not miss any malignant nodules for recommending follow-up and achieved specificity 34 of 84 (40%; 95% CI: 30%, 51%). Similar specificity ( $P = .74$ ) was obtained by the deep learning algorithm (specificity, 32 of 84; 38%; 95% CI: 28%, 49%). For the remaining nine readers, the mean sensitivity was 97%, whereas the mean specificity was relatively low (34%). In Table 2, we provide sensitivity and specificity for follow-up recommendation by all readers.

We split the test nodules that were positive for malignancy and negative for malignancy (ie, benign) into two subsets, easy and difficult, on the basis of the performance of human raters. Ten of 15 nodules positive for malignancy were included in the easy set on the basis of unanimous correct management decisions from all 10 readers (expert consensus and nine individual radiologists). For nodules that were negative for malignancy, 39 of 84 were also included in the easy set on the basis of at least six of 10 correct management decisions for FNA recommendation. These selections resulted in two subsets, one with 49 easy nodules (10 nodules positive for malignancy and 39 nodules negative for malignancy) and the other with 50 difficult nodules (five nodules positive for malignancy and 45 nodules negative for malignancy). Figure 6 compares the performance of deep learning and radiologists on a subset of easy (Fig 6a) and difficult (Fig 6b) test nodules. Deep learning achieved higher AUC than radiologists for the difficult nodules (0.92 vs 0.70, respectively;  $P = .02$ ) and similar AUC for the easy nodules (0.89 vs 0.92, respectively;  $P = .59$ ). Expert consensus and deep learning performed similarly for the difficult nodules (AUC, 0.90 [95% CI: 0.72, 1.00] vs 0.92 [95% CI: 0.80, 1.00], respectively;  $P = .96$ ). However, for the easy nodules, the deep learning AUC (0.89; 95% CI: 0.75, 0.98) was slightly lower than for expert consensus (0.96; 95% CI: 0.89, 0.99;  $P = .16$ ).

## Discussion

Interpretation of nodules at thyroid US is time consuming and has interreader variability. In our study, we developed a deep learning algorithm to provide management recommendations for thyroid nodules observed on US images and compared its performance with radiologists who adhered to American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) guidelines. We showed that the performance

**Table 1: Population Statistics according to Malignant Nodule Class**

Parameter	All Nodules ( <i>n</i> = 1377)	Training Nodules ( <i>n</i> = 1278)	Test Nodules ( <i>n</i> = 99)
Mean age of patient (y)	53.2 ± 14.0	53.2 ± 13.9	52.3 ± 14.0
Mean nodule size (cm)	2.6 ± 1.5	2.6 ± 1.5	2.7 ± 1.3
No. of malignant nodules	142 (10.3)	127 (9.9)	15 (15)

Note.—Data in parentheses are percentages; mean data are ± standard deviation.

of the algorithm was similar to that of consensus of three expert readers by achieving sensitivity of 87% (95% confidence interval [CI]: 67%, 100%) and specificity of 52% (95% CI: 41%, 63%).

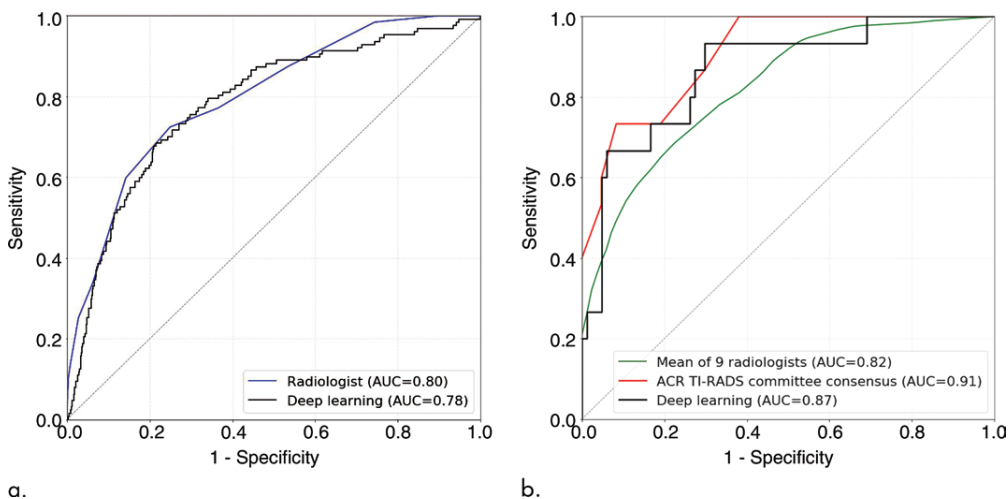
The most valuable aspect of the deep learning algorithm is the ability to improve specificity of thyroid nodule biopsy recommendations. In a study that compared the recommendations of eight radiologists for 100 nodules, Hoang et al (5) found that ACR TI-RADS offered a meaningful reduction in the number of thyroid nodules recommended for biopsy and improved specificity. In our study, we show that deep learning maintains or provides improvement in specificity compared with radiologists who use ACR TI-RADS, which suggests that the proposed algorithm offers performance markedly higher than radiologists who do not use ACR TI-RADS.

Our results add to the growing body of evidence demonstrating the potential power of deep learning when applied to thyroid US. Chi et al (14) showed that a system that uses imaging features extracted with a deep convolutional neural network can achieve accuracy greater than 99% for the binary task of classifying thyroid nodules on US images to ACR TI-RADS categories 1 and 2 versus all categories. Even though the performance seems to be outstanding, it refers to a greatly simplified task of predicting proxy labels. However, our ground truth used for both the training and testing nodule subsets relied on cytologic and pathologic results. In another study, Ma et al (15) used a large data set of over 8000 thyroid nodules with malignant and benign status confirmed either by operation or FNA result. The proposed deep learning algorithm that required manual nodule segmentation resulted in high sensitivity (82%) and specificity (84%); however, nodule sizes were not considered in the evaluation. The malignancy rate was also high in that study (15) and not reflective of a typical cohort of thyroid nodules undergoing thyroid US or biopsy. However, our study compared fully autonomous decisions made by a deep learning algorithm to radiologists.

A deep learning algorithm for prediction of malignancy could make a difference in clinical practice. First, for a given image, our algorithm will always provide the same prediction. Therefore, it will eliminate a substantial interreader variability that has been observed for this task even when the ACR TI-RADS system is used. Second, the algorithm could reduce the time required for interpretation of thyroid nodules, which puts some strain on radiology departments. Finally, deep learning may perform better than some radiologists who interpret thyroid US images in clinical practice, although a larger study is needed to confirm this.

The ACR TI-RADS system consists of two steps. The first step, on the basis of specific features of the nodules, estimates the likelihood that the lesion is malignant. The second step triages

nodules for biopsy or follow-up on the basis of the likelihood estimated in the first step and nodule size. Our deep learning system replaces only the first step and uses the same size-based triaging in the second stage. Whereas this design decision was important to allow for a fair comparison of our system with ACR TI-RADS in the proper clinical setting, to some extent it limits the system to the decision-making framework of ACR TI-RADS. Future improvement that considers the interactions between tumor size and more detailed features of the nodules could provide additional gains in performance in terms of sensitivity and specificity.



**Figure 5:** Areas under the receiver operating characteristic curves (AUCs) of **(a)** deep learning evaluated by using 10-fold cross-validation for 1278 training nodules compared with a single radiologist who used the American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) and **(b)** deep learning evaluated for 99 test nodules compared with expert consensus of three ACR TI-RADS committee members and nine radiologists who used ACR TI-RADS.

Our study had limitations. Our final test set of 99 nodules (15 nodules positive for malignancy and 84 nodules negative for malignancy) as well as easy and difficult test subsets contained a small number of nodules, which resulted in wide CIs. This limitation was alleviated by a cross-validation experiment on the larger training set (127 nodules positive for malignancy and 1151 nodules negative for malignancy), which showed results that were consistent with those from the test set in terms of the comparable performance of our algorithm with the radiologist who had the highest performance. Another limitation was that we noticed some differences in performance between the test set and the training set. This was not an indication of a high-bias model (ie, underfitting) because it was the case for both deep learning and the radiologist. We believe that the main reason for this difference is that the nodules from the training set were on average more difficult to interpret, which was corroborated by additional exploration of the data including evaluation of the discriminative power of features. Whereas the overall performance of all predictors (deep learning and radiologists) differed

between the test set and the training set. This was not an indication of a high-bias model (ie, underfitting) because it was the case for both deep learning and the radiologist. We believe that the main reason for this difference is that the nodules from the training set were on average more difficult to interpret, which was corroborated by additional exploration of the data including evaluation of the discriminative power of features. Whereas the overall performance of all predictors (deep learning and radiologists) differed

**Table 2: Comparison of the Deep Learning Algorithm, ACR TI-RADS Committee Expert Readers, and Radiologists**

Reader	FNA		Follow-up		AUC	Experience (y)
	Sensitivity	Specificity	Sensitivity	Specificity		
Deep learning algorithm	13/15 (87) [67, 100]	44/84 (52) [42, 62]	14/15 (93) [79, 100]	32/84 (38) [28, 49]	0.87 [0.76, 0.95]	NA
ACR TI-RADS committee expert readers ( $n = 3$ )	13/15 (87)	43/84 (51)	15/15 (100)	34/84 (40)	0.91	26–32
Radiologists ( $n = 9$ )						
Reader 1	14/15 (93)	40/84 (48)	15/15 (100)	28/84 (33)	0.91	20–25
Reader 2	13/15 (87)	24/84 (29)	15/15 (100)	14/84 (17)	0.76	20
Reader 3	12/15 (80)	40/84 (48)	15/15 (100)	27/84 (32)	0.85	13
Reader 4	12/15 (80)	40/84 (48)	15/15 (100)	28/84 (33)	0.83	13
Reader 5	11/15 (73)	49/84 (57)	14/15 (93)	34/84 (40)	0.78	3
Reader 6	11/15 (73)	59/84 (70)	13/15 (87)	51/84 (61)	0.85	32
Reader 7	12/15 (80)	42/84 (50)	15/15 (100)	33/84 (39)	0.81	4
Reader 8	13/15 (87)	32/84 (38)	14/15 (93)	19/84 (23)	0.79	32
Reader 9	14/15 (93)	37/84 (44)	15/15 (100)	26/84 (31)	0.83	20
Mean values for readers 1–9 (%)	83 [64, 98]	48 [37, 59]	97 [90, 100]	34 [24, 46]	0.82 [0.73, 0.90]	17

Note.—Unless otherwise indicated, data are numerator/denominator, data in parentheses are percentages, and data in brackets are 95% confidence intervals. The readers used the test set of 99 nodules. ACR = American College of Radiology, AUC = area under the receiver operating characteristic curve, FNA = fine-needle aspiration, NA = not applicable, TI-RADS = Thyroid Imaging Reporting and Data System.

between the two sets, the relative trends between radiologists and our algorithm remained. Regarding the study population, all nodules used in our study underwent FNA because of findings suspicious for malignancy or US findings that were indeterminate, and not on the basis of ACR TI-RADS guidelines. In addition, no large-scale test set from external institutions was available for comparison and to assess for generalization to a broader population of patients and nodules.

In summary, deep learning algorithms may be promising tools in the decision-making process for assessment of thyroid nodules. More studies are needed to further validate our findings.

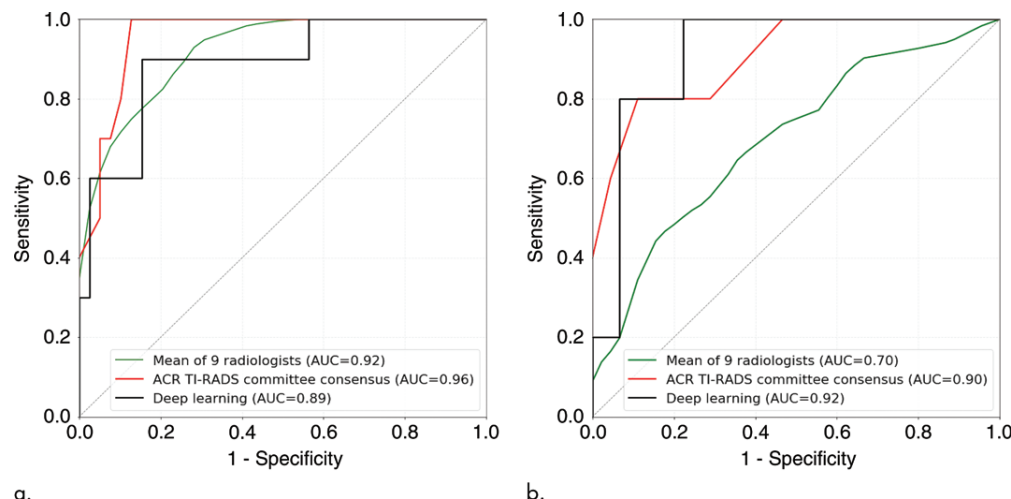
**Acknowledgments:** We thank Fernando J. Boschini, MD, Nirvikar Dahiya, MD, Jill E. Langer, MD, Justin R. Newman, MD, Carl C. Reading, MD, Daniel R. Scanga, MD, Sharlene A. Teefey, MD, Robert C. Vogler, MD, and four other radiologists who interpreted the test set of thyroid nodules as part of previously published work.

**Author contributions:** Guarantors of integrity of entire study, M.B., M.A.M.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, M.B., B.W.T., J.K.H., W.D.M.; clinical studies, D.T., W.D.M.; experimental studies, M.B., B.W.T., M.A.M.; statistical analysis, M.B., D.T., M.A.M.; and manuscript editing, all authors

**Disclosures of Conflicts of Interest:** M.B. disclosed no relevant relationships. B.W.T. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed an invention disclosure to the Duke University Office of Technology and Licensing. Other relationships: disclosed no relevant relationships. J.K.H. disclosed no relevant relationships. D.T. disclosed no relevant relationships. F.N.T. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed money paid to author for expert testimony from Starnes, Davis, Florie; disclosed speaking honoraria from the American College of Radiology. Other relationships: disclosed no relevant relationships. W.D.M. disclosed no relevant relationships. M.A.M. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed an invention disclosure to the Duke University Office of Technology and Licensing as well as advising relationship with Gradient Health. Other relationships: disclosed no relevant relationships.

## References

- Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid* 2016;26(1):1–133.
- Hoang JK, Middleton WD, Farjat AE, et al. Interobserver Variability of Sonographic Features Used in the American College of Radiology Thyroid Imaging Reporting and Data System. *AJR Am J Roentgenol* 2018;211(1):162–167.
- Vaccarella S, Franceschi S, Bray F, Wild CP, Plummer M, Dal Maso L. Worldwide Thyroid-Cancer Epidemic? The Increasing Impact of Overdiagnosis. *N Engl J Med* 2016;375(7):614–617.
- Tessler FN, Middleton WD, Grant EG, et al. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. *J Am Coll Radiol* 2017;14(5):587–595.



**Figure 6:** Areas under the receiver operating characteristic curves (AUCs) comparing deep learning, American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) committee consensus, and radiologists for (a) 49 easy test nodules and (b) 50 difficult test nodules. The easy nodules (a) include 10 malignant nodules on the basis of unanimous correct management decisions from nine readers and expert consensus and 39 benign nodules on the basis of at least six of 10 correct management decisions for fine-needle aspiration recommendation. The difficult nodules (b) include the remaining five malignant and 45 benign test nodules.

- Hoang JK, Middleton WD, Farjat AE, et al. Reduction in thyroid nodule biopsies and improved accuracy with American college of radiology thyroid imaging reporting and data system. *Radiology* 2018;287(1):185–193.
- Griffin AS, Mitsky J, Rawal U, Bronner AJ, Tessler FN, Hoang JK. Improved Quality of Thyroid Ultrasound Reports After Implementation of the ACR Thyroid Imaging Reporting and Data System Nodule Lexicon and Risk Stratification System. *J Am Coll Radiol* 2018;15(5):743–748.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436–444.
- Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115–118 [Published correction appears in *Nature* 2017;546(7660):686].
- Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316(22):2402–2410.
- Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. *RadioGraphics* 2017;37(2):505–515.
- Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. *J Digit Imaging* 2017;30(4):427–441.
- Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging* 2019;49(4):939–954.
- Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst* 2015; 91–99. <https://dl.acm.org/citation.cfm?id=2969250>.
- Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network. *J Digit Imaging* 2017;30(4):477–486.
- Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. *Ultrasonics* 2017;73:221–230.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Conf Comput Vis Pattern Recognit*, 2016, 770–778.
- Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. *Eur Conf Comput Vis*, 2014; 740–755.
- Buades A, Coll B, Morel JM. A non-local algorithm for image denoising. *Comput Vis Pattern Recognition*, 2005 CVPR 2005 IEEE Comput Soc Conf, 2005; 60–65.
- Coupé P, Hellier P, Kervrann C, Barillot C. Nonlocal means-based speckle filtering for ultrasound images. *IEEE Trans Image Process* 2009;18(10):2221–2229.
- Caruana R. Multitask learning. In: Thrun S, Pratt L, eds. *Learning to Learn*. Boston, Mass: Springer, 1998; 95–133.
- Nitish S, Hinton GE, Alex K, Ilya S Sr. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;15(1):1929–1958. <http://jmlr.org/papers/v15/srivastava14a.html>.
- Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. *arXiv Prepr arXiv:1708.02002*. <https://arxiv.org/abs/1708.02002>. Published August 7, 2017. Accessed DATE.
- Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw* 2018;106:249–259.

**A.7 Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility**

# Using Artificial Intelligence to Revise ACR TI-RADS Risk Stratification of Thyroid Nodules: Diagnostic Accuracy and Utility

Benjamin Wildman-Tobriner, MD • Mateusz Buda, MSc • Jenny K. Hoang, MBBS, MHS • William D. Middleton, MD • David Thayer, MD • Ryan G. Short, MD • Franklin N. Tessler, MD, CM • Maciej A. Mazurowski, PhD

From the Department of Radiology, Duke University Hospital, 2301 Erwin Rd, Durham, NC 27701 (B.W.T., M.B., J.K.H., R.G.S., M.A.M.); Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, Mo (W.D.M., D.T.); and Department of Radiology, University of Alabama at Birmingham, Birmingham, Ala (F.N.T.). Received September 14, 2018; revision requested November 12; final revision received March 7, 2019; accepted April 3. Address correspondence to B.W.T. (e-mail: benjamin.wildman-tobriner@duke.edu).

Conflicts of interest are listed at the end of this article.

Radiology 2019; 292:112–119 • <https://doi.org/10.1148/radiol.2019182128> • Content code: **HN**

**Background:** Risk stratification systems for thyroid nodules are often complicated and affected by low specificity. Continual improvement of these systems is necessary to reduce the number of unnecessary thyroid biopsies.

**Purpose:** To use artificial intelligence (AI) to optimize the American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS).

**Materials and Methods:** A total of 1425 biopsy-proven thyroid nodules from 1264 consecutive patients (1026 women; mean age, 52.9 years [range, 18–93 years]) were evaluated retrospectively. Expert readers assigned points based on five ACR TI-RADS categories (composition, echogenicity, shape, margin, echogenic foci), and a genetic AI algorithm was applied to a training set (1325 nodules). Point and pathologic data were used to create an optimized scoring system (hereafter, AI TI-RADS). Performance of the systems was compared by using a test set of the final 100 nodules with interpretations from the expert reader, eight nonexpert readers, and an expert panel. Initial performance of AI TI-RADS was calculated by using a test for differences between binomial proportions. Additional comparisons across readers were conducted by using bootstrapping; diagnostic performance was assessed by using area under the receiver operating curve.

**Results:** AI TI-RADS assigned new point values for eight ACR TI-RADS features. Six features were assigned zero points, which simplified categorization. By using expert reader data, the diagnostic performance of ACR TI-RADS and AI TI-RADS was area under the receiver operating curve of 0.91 and 0.93, respectively. For the same expert, specificity of AI TI-RADS (65%, 55 of 85) was higher ( $P < .001$ ) than that of ACR TI-RADS (47%, 40 of 85). For the eight nonexpert radiologists, mean specificity for AI TI-RADS (55%) was also higher ( $P < .001$ ) than that of ACR TI-RADS (48%). An interactive AI TI-RADS calculator can be viewed at <http://deckard.duhs.duke.edu/~ai-ti-rads>.

**Conclusion:** An artificial intelligence–optimized Thyroid Imaging Reporting and Data System (TI-RADS) validates the American College of Radiology TI-RADS while slightly improving specificity and maintaining sensitivity. Additionally, it simplifies feature assignments, which may improve ease of use.

© RSNA, 2019

Online supplemental material is available for this article.

Thyroid nodules are an extremely common finding at US and other imaging studies (1,2). Although most thyroid nodules are benign, many patients are subjected to a costly workup that may include one or more biopsies, follow-up imaging, and even diagnostic lobectomy (3). This contributes to the overdiagnosis of thyroid cancers that are not clinically significant (4). Over the past decade, multiple groups have developed biopsy guidelines for thyroid nodules based on their appearance at US, but some guidelines are difficult to apply and all lead to high false-positive rates (benign nodules for which biopsy is recommended).

With these issues in mind, a committee of the American College of Radiology (ACR) created the Thyroid Imaging Reporting and Data System (TI-RADS) to determine if thyroid nodules depicted at US require biopsy or follow-up (5). Nodules are awarded points based on features in five

categories—composition, echogenicity, shape, margin, and echogenic foci. The more suspicious the feature, the higher its point value. Points are summed to categorize a nodule into one of five TI-RADS risk levels, TR1 to TR5 (Table 1). Management recommendations are determined by using the risk level and the maximum size of the nodule.

The points assigned to each feature in ACR TI-RADS were based on evidence in the literature and expert consensus. Therefore, it is possible that the performance of the system could be improved by optimization of the points assigned to each US feature. Given the problem of overdiagnosis in thyroid imaging, this might improve specificity without sacrificing sensitivity. Indeed, the ACR TI-RADS committee recognized that certain features may warrant higher or lower point values to achieve optimal performance (5).



## Abbreviations

ACR = American College of Radiology, AI = artificial intelligence, CI = confidence interval, TI-RADS = Thyroid Imaging Reporting and Data System

## Summary

Artificial intelligence modeling suggests that the American College of Radiology Thyroid Imaging Reporting and Data System may be modified to improve ease of use while also improving specificity.

## Key Points

- By using a set of 1425 thyroid nodules, artificial intelligence (AI) modeling was used to optimize the American College of Radiology Thyroid Imaging Reporting and Data System (TI-RADS).
- The revised TI-RADS (hereafter, AI TI-RADS) assigned new point values for eight features, including a simplified scheme for some categories. For example, only assigning points to solid nodules and eliminating point assignments to other composition features represents one such modification.
- AI TI-RADS resulted in slightly higher specificity for recommending fine-needle aspiration (mean increase of 7.6% across eight radiologist readers;  $P < .001$ ).

The aim of this study was to use artificial intelligence (AI) algorithms to optimize TI-RADS feature point assignments. Our hypothesis was that our algorithm (hereafter, AI TI-RADS) could achieve similar or higher specificity than could ACR TI-RADS while maintaining sensitivity. This hypothesis would be tested in part by using a set of 100 thyroid nodules that had been interpreted by multiple radiologists as part of another study (although outcomes for our study would be different and use separate data analysis) (6). These results could both validate the current ACR TI-RADS and inform future revisions of the system.

## Materials and Methods

### Study Population and Image Annotation

This retrospective study was Health Insurance Portability and Accountability Act compliant, institutional review board approved, and used patients from a single academic medical center. A waiver of consent was obtained due to the anonymous and retrospective nature of the study. The initial population included 1631 thyroid nodules in 1439 consecutive patients who underwent diagnostic thyroid US and subsequent biopsy between August 2006 and May 2010. Sonograms were performed for a variety of clinical indications by using commercially available units (Antares and Elegra [Siemens Healthineers, Erlangen, Germany], ATL HDI 5000 and iU22 [Philips, Best, the Netherlands], and Logiq E9 [General Electric, Andover, Mass]). All were considered high-end units at the time that the images were obtained.

Tissue samples were obtained by using standard fine-needle aspiration techniques, and the cytopathologic slides were reviewed by pathology faculty at the institution where the images were obtained. Diagnosis was based on fine-needle aspiration results and surgical specimens, when available. Only nodules that were malignant or benign were included, unless a nodule underwent repeat fine-needle aspiration or surgical resection that confirmed

**Table 1: Risk Categories of American College of Radiology Thyroid Imaging Reporting and Data System**

Risk Category	Recommendation
TR1	Benign: no FNA
TR2	Not suspicious: no FNA
TR3	Mildly suspicious: FNA if $\geq 2.5$ cm
TR4	Moderately suspicious: FNA if $\geq 1.5$ cm
TR5	Highly suspicious: FNA if $\geq 1$ cm

Note.—FNA = fine-needle aspiration.

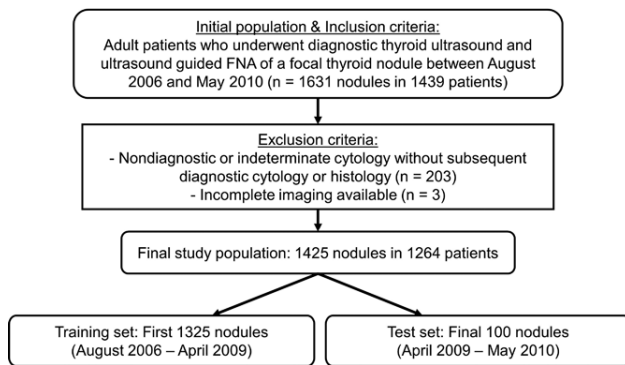
malignancy or benignity. Two hundred three nodules were excluded for indeterminate or nondiagnostic pathologic results and three were excluded because of incomplete images. The final study population comprised 1425 nodules from 1264 patients (Fig 1) with 151 (10.6%) cancers (95 papillary, 40 follicular variants of papillary, six follicular, one medullary, and nine other cancers).

Sonograms were interpreted by one of two expert readers (reader 1 [W.D.M.], with 20 years of experience and reader 2, with 20 years of experience; both members of ACR TI-RADS committee) who were blinded to the indication and pathologic result. Readers were not blinded to patient age, as this was included on the images. Reader 1 interpreted 1044 (64%) of the nodules and reader 2 interpreted 587 (36%). They jointly interpreted 50 cases at the beginning of the study to standardize their approach and read another 50 cases together in the middle of the study. The expert readers assigned features in the five ACR TI-RADS categories for every nodule. Because this was performed prior to the publication of ACR TI-RADS, there were minor differences in terminology for echogenicity and shape. Hypoechoic nodules were characterized as mildly, moderately, or very hypoechoic rather than hypoechoic or very hypoechoic as in ACR TI-RADS. Therefore, nodules that were originally called mildly hypoechoic were recategorized as hypoechoic, and nodules originally categorized as moderately to very hypoechoic were reclassified as hypoechoic or very hypoechoic by a third reader (reader 3 [B.W.T.], a radiology fellow with 5 years daily practice in thyroid imaging). Nodule shape was also determined by reader 3 because this feature was not part of the original analysis. Recategorizations that were deemed difficult or indeterminate were reviewed by a fourth reader (reader 4 [J.K.H.], with 13 years of experience in thyroid imaging; member of ACR TI-RADS committee). In all other respects, the original analysis followed the recommendations of ACR TI-RADS.

Ultimately, all 1425 nodules had feature assignments for all five ACR TI-RADS categories, which yielded point assignments and corresponding TI-RADS risk levels. Nodules were split into a training set of 1325 nodules (1189 benign, 136 malignant) and a test set of the last 100 nodules (85 benign, 15 malignant). A validation set was not used; rather, cross-validation within the training cases was used to tune the algorithm.

### AI TI-RADS Algorithm Development

We used a genetic algorithm to derive an optimized and data-driven version of TI-RADS, which we refer to as AI TI-RADS



**Figure 1:** Flowchart illustrates exclusion criteria and final study population. FNA = fine-needle aspiration.

(7). Genetic algorithms are a part of computational intelligence methods, a subgroup of AI methods that focus on algorithms inspired by natural selection and its genetic underpinnings. Specifically, a population of individuals is simulated by a computer algorithm in which each individual represents a solution to a problem. In this instance, the solution was a set of points for different thyroid nodule features. Each individual (representing a possible solution) was evaluated in terms of its “fitness,” which reflected how accurately the set of points could predict malignancy. Through multiple iterations (“generations”), individuals with better performance were prioritized and multiplied. This process was repeated 50 times, and eventually a single best solution was presented. This optimized set of AI TI-RADS points had the same form as did the original ACR TI-RADS, but with different point values for some features. Therefore, the proposed system could immediately be used in the same manner as ACR TI-RADS. Additional details of the genetic algorithm are presented in Appendix E1 (online) and the following link can be accessed for additional code details: <https://github.com/mateuszbuda/AI-TI-RADS/releases/tag/v1.0>.

### Comparing AI TI-RADS with ACR TI-RADS

After new point values were assigned to some TI-RADS features, the two systems were applied to the test set as interpreted by the expert reader. The test set was also interpreted by 11 other radiologists as part of another previously published study (6). Three radiologists (F.N.T., with 34 years of experience in thyroid imaging; reader 5, with 26 years of experience in thyroid imaging; and reader 6, with 31 years of experience in thyroid imaging; all members of the ACR TI-RADS committee) independently interpreted the 100 nodules and their consensus was taken as the best possible performance for the test set. The eight other radiologist readers (two academic, six general private practice [range, 3–32 years of experience in thyroid imaging]) had not routinely used ACR TI-RADS at the time of interpretation. After initial training, they assigned features to each nodule according to ACR TI-RADS, and points and risk categories were assigned by using both systems.

### Statistical Analysis

By using the single-expert reader data, the area under the receiver operating curve, sensitivity, and specificity were calcu-

lated for ACR TI-RADS and AI TI-RADS models by using a test for differences between two binomial proportions. Sensitivity and specificity for detection of malignancy were also calculated for each nonexpert reader and the expert consensus, and comparison across those groups was performed by using bootstrapping methods. Differences in mean age between men and women were tested by using an unpaired *t* test. Statistical analysis was conducted by using R software (R Foundation for Statistical Computing, Vienna, Austria; <https://r-project.org>), and *P* values less than or equal to .05 were considered to indicate statistical significance.

## Results

### Study Population and Nodule Characteristics

The mean of the summed ACR TI-RADS points was  $4.23 \pm 2.45$  (standard deviation) for the 1325 training nodules and  $4.22 \pm 2.64$  for the 100 test nodules. The mean age for male patients was 56.7 years  $\pm 13.5$ , whereas the mean age for female patients was 52.0 years  $\pm 13.9$  ( $P < .001$ ). Additional basic demographics for the training and test sets can be found in Table 2. The distribution for all TI-RADS imaging features across all included nodules for the training and test sets appears in Table 3.

### AI TI-RADS Algorithm

Figure 2 displays the AI TI-RADS and ACR TI-RADS point values. AI TI-RADS differed slightly from ACR TI-RADS in all five feature categories. The AI algorithm assigned new point values for eight features. Six of the eight features with new values changed by one point, while the other two features (taller than wide for shape and “cannot tell” for composition) each changed by two points. The overall order of features within each category was preserved. The highest risk features maintained the greatest point values. Figure 3 shows the final classification system including size cutoffs and management recommendations. An interactive AI TI-RADS calculator can be viewed at <http://deckard.duhs.duke.edu/ai-ti-rads>.

For composition, AI TI-RADS assigned three points for solid or almost solid nodules and no points to the three other features under this category (as well as no points to “cannot tell”). For echogenicity, points for hypoechoic and very hypoechoic were the same under both systems, but AI TI-RADS assigned no points for other features within the category. For shape, AI TI-RADS assigned one point for taller-than-wide shape compared with three points for ACR TI-RADS. For margin, an irregular and/or lobulated margin was assigned the same number of points in each system. For echogenic foci, AI TI-RADS assigned no points to macrocalcifications compared with one point in ACR TI-RADS. The other feature point assignments under the echogenic foci category were the same for both systems.

### Comparing AI TI-RADS with ACR TI-RADS

When both systems were applied to the test set of 100 nodules, AI TI-RADS assigned lower TI-RADS risk levels than did ACR TI-RADS for 43 nodules. Specifically, five nodules were downgraded from TR5 to TR4, 11 nodules were changed from TR4

**Table 2: Patient Demographics and Nodule Characteristics in the Training and Test Sets**

Demographic	Training Benign ( <i>n</i> = 1189)	Training Malignant ( <i>n</i> = 136)	Test Benign ( <i>n</i> = 85)	Test Malignant ( <i>n</i> = 15)
Sex*				
Female	874 (81.9)	99 (78.0)	62 (77.5)	10 (71.4)
Male	186 (17.4)	27 (21.3)	17 (21.3)	3 (21.4)
Unknown	7 (0.7)	1 (0.8)	1 (1.3)	1 (7.1)
Age (y) <sup>†</sup>	53.4 ± 13.7 (18–93)	50.1 ± 15.5 (21–89)	53.7 ± 13.7 (26–82)	46.1 ± 13.2 (19–68)
Female	52.6 ± 13.6 (18–93)	49.0 ± 16.1 (21–89)	52.7 ± 13.6 (26–82)	42.7 ± 13.5 (19–62)
Male	56.8 ± 13.7 (18–82)	55.1 ± 11.7 (25–78)	58.9 ± 12.1 (31–76)	56.0 ± 8.5 (49–68)
Mean nodule size (cm) <sup>‡</sup>	2.7 ± 1.5	2.0 ± 1.3	2.8 ± 1.3	2.2 ± 1.2
ACR TI-RADS risk level				
TR1	122 (10.3)	0 (0.0)	11 (12.9)	0 (0.0)
TR2	181 (15.2)	2 (1.5)	14 (16.5)	0 (0.0)
TR3	250 (21.0)	16 (11.8)	19 (22.4)	0 (0.0)
TR4	468 (39.4)	40 (29.4)	32 (37.6)	3 (20.0)
TR5	168 (14.1)	78 (57.4)	9 (10.6)	12 (80.0)

Note.—Unless otherwise specified, data are numbers of nodules, with percentages in parentheses. ACR = American College of Radiology, TI-RADS = Thyroid Imaging Reporting and Data System.

\* Denotes number of patients within each category. Numbers in parentheses represent percentage within a given group (benign, malignant).

<sup>†</sup> Data are means ± standard deviation, with ranges in parentheses.

<sup>‡</sup> Data are means ± standard deviation.

**Table 3: Distribution of TI-RADS Features across the Training and Test Sets**

Feature	Training Benign ( <i>n</i> = 1189)	Training Malignant ( <i>n</i> = 136)	Test Benign ( <i>n</i> = 85)	Test Malignant ( <i>n</i> = 15)
Composition				
Cystic or almost completely cystic	3 (0.3)	0 (0.0)	0 (0.0)	0 (0.0)
Spongiform	119 (10.0)	0 (0.0)	11 (12.9)	0 (0.0)
Mixed cystic and solid	480 (40.4)	20 (14.7)	39 (45.9)	0 (0.0)
Solid or almost completely solid	580 (48.8)	116 (85.3)	35 (41.2)	15 (100)
Cannot tell	7 (0.6)	0 (0.0)	0 (0.0)	0 (0.0)
Echogenicity				
Anechoic	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Hyperechoic	146 (12.3)	7 (5.1)	20 (23.5)	2 (13.3)
Isoechoic	473 (39.8)	22 (16.2)	31 (36.5)	3 (20.0)
Hypoechoic	497 (41.8)	89 (65.4)	33 (38.8)	7 (46.7)
Very hypoechoic	38 (3.2)	15 (11.0)	1 (1.2)	3 (20.0)
Cannot classify	35 (2.9)	3 (2.2)	0 (0.0)	0 (0.0)
Shape				
Taller than wide	119 (10.0)	25 (18.4)	9 (10.6)	2 (13.3)
Not taller than wide	1070 (90.0)	111 (81.6)	76 (89.4)	13 (86.7)
Margin				
Smooth	830 (69.8)	75 (55.1)	64 (75.3)	5 (33.3)
Ill-defined	272 (22.9)	25 (18.4)	14 (16.5)	1 (6.7)
Irregular and/or lobulated	81 (6.8)	34 (25.0)	4 (4.7)	9 (60.0)
Extrathyroidal extension	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Cannot classify	6 (0.5)	2 (1.5)	3 (3.5)	0 (0.0)
Echogenic foci				
No echogenic foci	714 (60.1)	44 (32.4)	53 (62.4)	3 (20.0)
Large comet-tail artifacts	87 (7.3)	2 (1.5)	7 (8.2)	1 (6.7)
Macrocalcifications	167 (14.0)	26 (19.1)	8 (9.4)	3 (20.0)
Peripheral calcifications	42 (3.5)	11 (8.1)	0 (0.0)	1 (6.7)
Punctate echogenic foci	266 (22.4)	74 (54.4)	22 (25.9)	10 (66.7)

Note.—Data are numbers of nodules, with percentages in parentheses. Nodules could have more than one type of echogenic focus. TI-RADS = Thyroid Imaging Reporting and Data System.

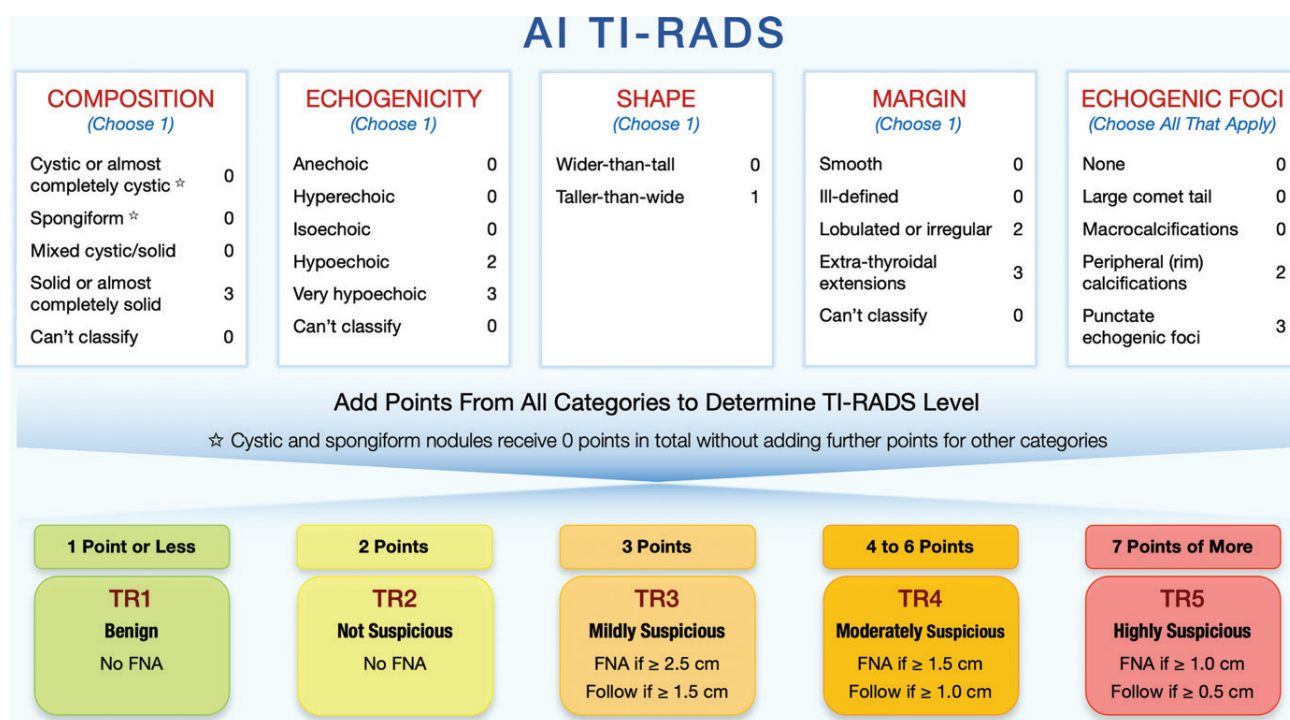
ACR TI-RADS	COMPOSITION (Choose 1)		ECHOGENICITY (Choose 1)		SHAPE (Choose 1)		MARGIN (Choose 1)		ECHOGENIC FOCI (Choose All That Apply)	
	Cystic or almost completely cystic	0	Anechoic	0	Wider-than-tall	0	Smooth	0	None	0
	Spongiform	0	Hyperechoic	1	Taller-than-wide	3	Ill-defined	0	Large comet tail	0
	Mixed cystic/solid	1	Isoechoic	1			Irregular/lobulated	2	Macrocalcifications	1
	Solid or almost completely solid	2	Hypoechoic	2			Extra-thyroidal extensions	3	Peripheral	2
	Can't classify	2	Very hypoechoic	3			Can't classify	0	Punctate	3
			Can't classify	1						

AI TI-RADS	COMPOSITION (Choose 1)		ECHOGENICITY (Choose 1)		SHAPE (Choose 1)		MARGIN (Choose 1)		ECHOGENIC FOCI (Choose All That Apply)	
	Cystic or almost completely cystic	0	Anechoic *	0	Wider-than-tall	0	Smooth	0	None	0
	Spongiform	0	Hyperechoic	0	Taller-than-wide	1	Ill-defined	0	Large comet tail	0
	Mixed cystic/solid	0	Isoechoic	0			Irregular/lobulated	2	Macrocalcifications	0
	Solid or almost completely solid	3	Hypoechoic	2			Extra-thyroidal extensions *	3	Peripheral	2
	Can't classify	0	Very hypoechoic	3			Can't classify	0	Punctate	3
			Can't classify	0						

\* Could not be evaluated due to small sample size. Points adapted from ACR TI-RADS.

**Figure 2:** Image shows comparison of American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) to Artificial Intelligence (AI) TI-RADS. Multiple new point assignments were designated by algorithm, including changing point values to zero for several features.

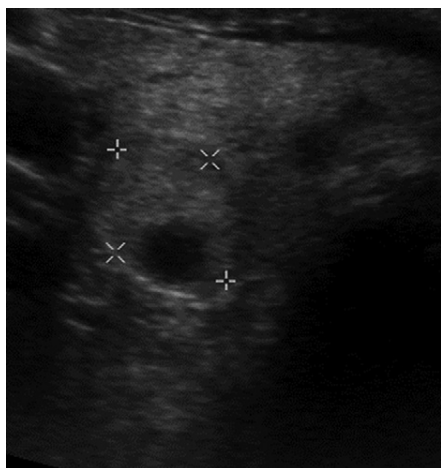


**Figure 3:** Image shows Artificial Intelligence (AI) Thyroid Imaging Reporting and Data System (TI-RADS) classification scheme, including nodule sizes that dictate follow-up recommendations. Nodule size cutoffs were kept the same as American College of Radiology TI-RADS. FNA = fine-needle aspiration, TR = TI-RADS category.

to TR3, three nodules were lowered from TR4 to TR2, two nodules were changed from TR4 to TR1, eight nodules were lowered from TR3 to TR2, and 14 nodules were reassigned from

TR2 to TR1. There were no nodules for which AI TI-RADS assigned a higher risk level than did ACR TI-RADS. Ultimately, the new risk level assignments resulted in 15 nodules for which



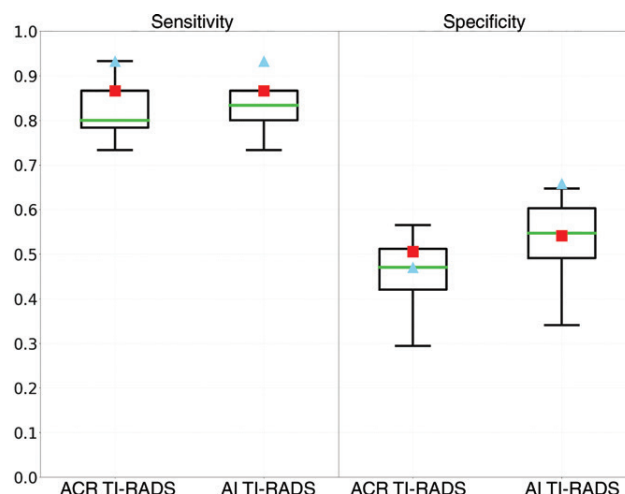


**Figure 4:** Image in a 72-year-old woman with right thyroid nodule. Transverse US image shows mixed cystic and solid, isoechoic, taller-than-wide nodule. Its features earn five total points according to American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) with risk level of TR4 and recommendation for fine-needle aspiration (FNA). Nodule earns only one point by using Artificial Intelligence (AI) TI-RADS with risk level of TR1 and no recommendation for FNA. Pathologic finding at FNA was benign nodule. Calipers were included in all images as part of automated nodule detection process.



**Figure 5:** Image in a 66-year-old man with right thyroid nodule. Long US image shows mixed cystic and solid hypoechoic nodule. This nodule earns three points according to American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) with risk level of TR3 and recommendation for biopsy. Nodule earns only two points with Artificial Intelligence (AI) TI-RADS with risk level of TR2 and no recommendation for fine-needle aspiration (FNA). FNA revealed benign thyroid nodule. Calipers were included in all images as part of automated nodule detection process.

ACR TI-RADS recommended fine-needle aspiration but AI TI-RADS did not, and all 15 nodules were benign (examples shown in Figs 4, 5). There were no nodules for which AI TI-RADS recommended fine-needle aspiration but ACR TI-RADS did not.



**Figure 6:** Box and whisker plot shows sensitivity and specificity for American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) compared with Artificial Intelligence (AI) TI-RADS. Boxes correspond to 25th and 75th percentiles for eight non-expert readers. Whiskers denote maximum and minimum values, and green line represents median. Red squares represent expert consensus, and blue triangles represent single expert reader. Sensitivity of both systems was similar, while specificity was higher for all groups when using AI TI-RADS.

By using the single-expert reader data, the area under the receiver operator curves for each system were similar: 0.91 (95% confidence interval [CI]: 0.82, 0.98) for ACR TI-RADS and 0.93 (95% CI: 0.85, 0.98) for AI TI-RADS ( $P = .18$ ). Their sensitivities (for detection of malignancy through recommendation of fine-needle aspiration) were the same (14 of 15, 93.3%; 95% CI: 77%, 100% for both), whereas the specificity of AI TI-RADS (55 of 85, 64.7%; 95% CI: 54%, 74%) was higher than was ACR TI-RADS (40 of 85, 47.0%; 95% CI: 37%, 57%) ( $P < .001$ ) (Fig 6).

When ACR TI-RADS and AI TI-RADS were applied to the test set interpretation of the eight nonexpert radiologists, AI TI-RADS had higher specificity than did ACR TI-RADS for every reader. Mean specificity of the eight readers by using AI TI-RADS was  $55.3\% \pm 12.8$  compared with  $47.5\% \pm 12.3$  by using ACR TI-RADS ( $P < .001$ ) (Table 4). The sensitivity of AI TI-RADS was lower for five of the eight nonexpert radiologists, although the mean sensitivity was not significantly lower (Table 4). Performance for the expert panel was similar, although the small increase in specificity was not statistically significant (Table 4).

## Discussion

Risk stratification systems for thyroid nodules at US are often affected by low specificity and poor interobserver agreement. We applied a machine learning technique to American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) to optimize the performance of the system while still maintaining the TI-RADS lexicon and structure. Our data-driven artificial intelligence (AI)-optimized version of TI-RADS (hereafter, AI TI-RADS) validates ACR TI-RADS: feature point allocations were the same for 15 of 23 features, and

**Table 4: Performance Comparison for Three Different Sets of Readers When Using ACR TI-RADS versus AI TI-RADS**

Reader	ACR TI-RADS		AI TI-RADS			
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	<i>P</i> Value	Specificity (%)	<i>P</i> Value
Single expert reader	14/15 (93.3) [77.2, 100]	40/85 (47.1) [37.3, 57.1]	14/15 (93.3) [77.2, 100]	NA	55/85 (64.7) [54.5, 74]	< .001
Mean of eight nonexpert readers*	81.7 (62.5, 97.7)	47.7 (36.4, 59.0)	82.5 (64.1, 97.7)	<i>P</i> > .5	55.3 (43.7, 66.6)	< .001
Expert panel consensus	13/15 (86.7) [66.7, 100.0]	43/85 (50.6) [40.3, 61.0]	13/15 (86.7) [66.7, 100.0]	NA	46/85 (54.1) [43.7, 64.6]	.10

Note.—Unless otherwise specified, data are numerators and denominators, with percentages in parentheses and 95% confidence intervals in brackets. *P* values reflect comparison of American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS) to Artificial Intelligence (AI) TI-RADS within a reader group.

\* Data in parentheses are 95% confidence intervals.

the highest risk features maintained the highest point values. However, alterations in point assignments under AI TI-RADS suggest that ACR TI-RADS may be simplified, as six features were assigned new point values of zero. Despite simplification, our results show a modest increase of 7% in mean specificity when applied to eight nonexpert radiologists.

ACR TI-RADS was based on literature review, expert consensus, and partial analysis of a database of proven nodules, and early studies of the system are encouraging. The system was validated in a multi-institutional study of more than 3400 nodules (8), and more recent retrospective studies have shown that it reduces nodule biopsy recommendations and improves accuracy compared with other biopsy guidelines (9–12). The point assignments derived from our AI TI-RADS model were similar to those of the ACR version, adding to the growing body of evidence supporting its use. Although point values in our AI model were different for eight features, most changed by only one point. Moreover, the area under the receiver operating curves for our data set by using ACR TI-RADS and AI TI-RADS were similar (0.91 and 0.93, respectively) and higher than that described in a recent analysis by Pantano et al (9) (area under the receiver operating curve, 0.78). Overall, our data support ACR TI-RADS.

ACR TI-RADS and our AI TI-RADS model had comparable receiver operating characteristic performance by using interpretations by two experts, although AI TI-RADS yielded slightly higher specificity and fewer recommendations for fine-needle aspiration. When applied to eight nonexpert readers, AI TI-RADS again had a small but statistically significant increase in specificity and minimal impact on sensitivity. This achieves a central aim of ACR TI-RADS: to focus on clinically significant thyroid cancers and to reduce fine-needle aspiration of benign nodules (5). It has been reported that overdiagnosis accounts for up to 77% of cases of thyroid cancer (4) and that more thyroid cancer diagnoses do not reduce mortality (13). Therefore, a small reduction in sensitivity seems acceptable in light of a larger gain in specificity. As well, many nodules not biopsied would meet the criteria for follow-up, mitigating the likelihood of missing cancers while potentially reducing health care costs.

Although AI TI-RADS had substantial overlap with and similar performance to ACR TI-RADS, the altered feature

values suggest that ACR TI-RADS may be simplified (Fig 2). For example, in the composition category, nodules are assigned four different possible point values in ACR TI-RADS, whereas our AI TI-RADS model assigned three points to solid nodules and zero points for all other types. This simplified scheme, which focuses on solid nodules, aligns with data from Middleton et al (8), who showed that solid nodules had four times higher risk of malignancy than did mixed cystic and solid nodules. Two meta-analyses (14,15) have also shown that solid composition confers some degree of risk, but they did not directly compare them to mixed cystic and solid nodules. This modification would allow a reader to focus on only one feature within the composition category and may improve efficiency.

AI TI-RADS point assignment in the echogenic foci category also differed. Peripheral calcifications and punctate echogenic foci were unchanged, but AI TI-RADS assigned zero points to macrocalcifications (compared with one point for ACR TI-RADS). This would simplify a category that already contains multiple features with low interobserver variability compared with other TI-RADS features (6). This modification highlights ongoing uncertainty regarding the clinical importance of macrocalcifications (16–18). Some studies (19) suggest that they confer a higher degree of risk than originally thought, whereas other studies (8) suggest that this feature is less suspicious than are punctate echogenic foci or peripheral calcifications. Although the allocation of zero points does not imply zero risk, this suggests that macrocalcifications are less suspicious than are punctate echogenic foci or peripheral calcifications based on our data.

The three remaining TI-RADS categories—echogenicity, shape, and margin—were also simplified by AI TI-RADS. The algorithm eliminated points for hyperechoic and isoechoic nodules in the echogenicity category but preserved two and three points for the higher-risk hypoechoic and very hypoechoic features. AI TI-RADS also reduced the number of points for taller-than-wide nodules from three to one, a result that contradicts studies that showed taller-than-wide shape as a high-risk and specific marker of malignancy (14,15,20). The reason for this is unclear, but may be related to low sample size.

Our study had some limitations. The training set was collected from a single institution and feature assignments were based on expert readers. Although there was potential for

overfitting given that an expert reader interpreted cases in both the training and test sets, the eight general radiologists were only interpreting the test set and performed relatively similarly to the expert, suggesting a reasonable fit for the model. In addition, a subset of features was assigned by a radiology fellow; however, cases were reviewed with an expert reader (member of the ACR TI-RADS committee), and as before, test data from the eight nonexpert readers and the expert consensus were not modified and helped to validate the model. Another limitation was the possibility of bias due to the test set being taken from the end of the study period, when newer scanners with improved image quality may have become available. However, scanners throughout the study period were considered high quality. We did not use a validation set as part of our training. Rather, we used cross-validation within the training cases. After all the hyperparameters were selected, we fixed them and trained by using entire training set. We chose this approach because of the relatively limited number of cases available. As well, features of extrathyroidal extension and cystic composition did not have enough data to be analyzed. However, they represent extremes of the risk spectrum; the decision to biopsy or not is clear when either of these findings are present. We used integers to make new point assignments in AI TI-RADS to mimic ACR TI-RADS and to simplify categorization. This resulted in some features earning zero points, which may falsely imply that a given feature confers no risk of malignancy. Because AI TI-RADS removed points for seven features and added points for only one feature, it is unsurprising that more nodules were reassigned to a lower TI-RADS level. This would be expected to improve specificity (which was the case), but it is somewhat surprising that there was not a corresponding decrease in sensitivity, possibly due to the relatively small number of malignant nodules (15) in the test set. There were also limitations related to our use of the computational models. For example, a high dimensional input space was relatively scarcely represented. Therefore, for unusual inputs, it is possible that the model returned unexpected and difficult-to-explain results.

Nonetheless, to our knowledge, this study represents a unique computational validation of one of the numerous risk stratification systems for thyroid nodules. Continued performance improvement is vital, and subsequent work could focus on further enhancements. Increasing the number of training cases represents a possible avenue for improvement. Future efforts could also include nodules with indeterminate pathologic results to broaden the mix of nodules included, which may enhance generalizability and performance.

In conclusion, artificial intelligence (AI)-optimized Thyroid Imaging Reporting and Data System (TI-RADS) is a data-driven model for risk stratification of thyroid nodules at US that both validates American College of Radiology (ACR) TI-RADS and suggests modifications to it that may improve its performance and enhance applicability. Prospective studies with long-term follow-up will be needed to refine ACR TI-RADS and assess its impact on clinical outcomes.

**Author contributions:** Guarantors of integrity of entire study, B.W.T., M.A.M.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, B.W.T., M.B., J.K.H., W.D.M.; clinical studies, W.D.M.; experimental studies, M.B., D.T., F.N.T., M.A.M.; statistical analysis, B.W.T., M.B., J.K.H., D.T., M.A.M.; and manuscript editing, all authors

**Disclosures of Conflicts of Interest:** B.W.T. disclosed no relevant relationships. M.B. disclosed no relevant relationships. J.K.H. disclosed no relevant relationships. W.D.M. disclosed no relevant relationships. D.T. disclosed no relevant relationships. R.G.S. disclosed no relevant relationships. F.N.T. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received payment for expert testimony from Davis, Florie, and Starnes; received payment from American College of Radiology for development of educational presentation and for travel/accommodations/meeting expenses unrelated to activities listed. Other relationships: disclosed no relevant relationships. M.A.M. disclosed no relevant relationships.

## References

- Hoang JK, Langer JE, Middleton WD, et al. Managing incidental thyroid nodules detected on imaging: white paper of the ACR Incidental Thyroid Findings Committee. *J Am Coll Radiol* 2015;12(2):143–150.
- Smith-Bindman R, Lebda P, Feldstein VA, et al. Risk of thyroid cancer based on thyroid ultrasound imaging characteristics: results of a population-based study. *JAMA Intern Med* 2013;173(19):1788–1796.
- Hobbs HA, Bahl M, Nelson RC, Eastwood JD, Escamado RM, Hoang JK. Applying the Society of Radiologists in Ultrasound recommendations for fine-needle aspiration of thyroid nodules: effect on workup and malignancy detection. *AJR Am J Roentgenol* 2014;202(3):602–607.
- Vaccarella S, Franceschi S, Bray F, Wild CP, Plummer M, Dal Maso L. Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. *N Engl J Med* 2016;375(7):614–617.
- Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR TI-RADS Committee. *J Am Coll Radiol* 2017;14(5):587–595.
- Hoang JK, Middleton WD, Farjat AE, et al. Interobserver variability of sonographic features used in the American College of Radiology Thyroid Imaging Reporting and Data System. *AJR Am J Roentgenol* 2018;211(1):162–167.
- Koza JR. Genetic programming as a means for programming computers by natural selection. *Stat Comput* 1994;4(2):87–112.
- Middleton WD, Teehey SA, Reading CC, et al. Multiinstitutional analysis of thyroid nodule risk stratification using the American College of Radiology Thyroid Imaging Reporting and Data System. *AJR Am J Roentgenol* 2017;208(6):1331–1341.
- Lauria Pantano A, Maddaloni E, Briganti SI, et al. Differences between ATA, AACE/ACE/AME and ACR TI-RADS ultrasound classifications performance in identifying cytological high-risk thyroid nodules. *Eur J Endocrinol* 2018;178(6):595–603.
- Zheng Y, Xu S, Kang H, Zhan W. A single-center retrospective validation study of the American College of Radiology Thyroid Imaging Reporting and Data System. *Ultrasound Q* 2018;34(2):77–83.
- Grani G, Lamartina L, Ascoli V, et al. Reducing the number of unnecessary thyroid biopsies while improving diagnostic accuracy: toward the “right” TIRADS. *J Clin Endocrinol Metab* 2019;104(1):95–102.
- Ha EJ, Na DG, Baek JH, Sung JY, Kim JH, Kang SY. US fine-needle aspiration biopsy for thyroid malignancy: diagnostic performance of seven society guidelines applied to 2000 thyroid nodules. *Radiology* 2018;287(3):893–900.
- Davies L, Welch HG. Current thyroid cancer trends in the United States. *JAMA Otolaryngol Head Neck Surg* 2014;140(4):317–322.
- Brito JP, Gionfriddo MR, Al Nofal A, et al. The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. *J Clin Endocrinol Metab* 2014;99(4):1253–1263.
- Remonti LR, Kramer CK, Leitão CB, Pinto LC, Gross JL. Thyroid ultrasound features and risk of carcinoma: a systematic review and meta-analysis of observational studies. *Thyroid* 2015;25(5):538–550.
- Lu Z, Mu Y, Zhu H, et al. Clinical value of using ultrasound to assess calcification patterns in thyroid nodules. *World J Surg* 2011;35(1):122–127.
- Na DG, Kim DS, Kim SJ, Ryoo JW, Jung SL. Thyroid nodules with isolated macrocalcification: malignancy risk and diagnostic efficacy of fine-needle aspiration and core needle biopsy. *Ultrasonography* 2016;35(3):212–219.
- Park YJ, Kim JA, Son EJ, et al. Thyroid nodules with macrocalcification: sonographic findings predictive of malignancy. *Yonsei Med J* 2014;55(2):339–344.
- Arpaci D, Ozdemir D, Cuhaci N, et al. Evaluation of cytopathological findings in thyroid nodules with macrocalcification: macrocalcification is not innocent as it seems. *Arq Bras Endocrinol Metabol* 2014;58(9):939–945.
- Moon WJ, Jung SL, Lee JH, et al. Benign and malignant thyroid nodules: US differentiation—multicenter retrospective study. *Radiology* 2008;247(3):762–770.

**A.8 A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images**





# A Data Set and Deep Learning Algorithm for the Detection of Masses and Architectural Distortions in Digital Breast Tomosynthesis Images

Mateusz Buda, MSc; Ashirbani Saha, PhD; Ruth Walsh, MD; Sujata Ghate, MD; Nianyi Li, PhD; Albert Świąćicki, BSc; Joseph Y. Lo, PhD; Maciej A. Mazurowski, PhD

## Abstract

**IMPORTANCE** Breast cancer screening is among the most common radiological tasks, with more than 39 million examinations performed each year. While it has been among the most studied medical imaging applications of artificial intelligence, the development and evaluation of algorithms are hindered by the lack of well-annotated, large-scale publicly available data sets.

**OBJECTIVES** To curate, annotate, and make publicly available a large-scale data set of digital breast tomosynthesis (DBT) images to facilitate the development and evaluation of artificial intelligence algorithms for breast cancer screening; to develop a baseline deep learning model for breast cancer detection; and to test this model using the data set to serve as a baseline for future research.

**DESIGN, SETTING, AND PARTICIPANTS** In this diagnostic study, 16 802 DBT examinations with at least 1 reconstruction view available, performed between August 26, 2014, and January 29, 2018, were obtained from Duke Health System and analyzed. From the initial cohort, examinations were divided into 4 groups and split into training and test sets for the development and evaluation of a deep learning model. Images with foreign objects or spot compression views were excluded. Data analysis was conducted from January 2018 to October 2020.

**EXPOSURES** Screening DBT.

**MAIN OUTCOMES AND MEASURES** The detection algorithm was evaluated with breast-based free-response receiver operating characteristic curve and sensitivity at 2 false positives per volume.

**RESULTS** The curated data set contained 22 032 reconstructed DBT volumes that belonged to 5610 studies from 5060 patients with a mean (SD) age of 55 (11) years and 5059 (100.0%) women. This included 4 groups of studies: (1) 5129 (91.4%) normal studies; (2) 280 (5.0%) actionable studies, for which where additional imaging was needed but no biopsy was performed; (3) 112 (2.0%) benign biopsied studies; and (4) 89 studies (1.6%) with cancer. Our data set included masses and architectural distortions that were annotated by 2 experienced radiologists. Our deep learning model reached breast-based sensitivity of 65% (39 of 60; 95% CI, 56%-74%) at 2 false positives per DBT volume on a test set of 460 examinations from 418 patients.

**CONCLUSIONS AND RELEVANCE** The large, diverse, and curated data set presented in this study could facilitate the development and evaluation of artificial intelligence algorithms for breast cancer screening by providing data for training as well as a common set of cases for model validation. The performance of the model developed in this study showed that the task remains challenging; its performance could serve as a baseline for future model development.

JAMA Network Open. 2021;4(8):e2119100. doi:10.1001/jamanetworkopen.2021.19100

**Open Access.** This is an open access article distributed under the terms of the CC-BY License.

JAMA Network Open. 2021;4(8):e2119100. doi:10.1001/jamanetworkopen.2021.19100

## Key Points

**Question** Can a curated, annotated, and publicly available data set of digital breast tomosynthesis (DBT) volumes be created for the development and validation of breast cancer computer-aided detection algorithms?

**Findings** In this diagnostic study, a curated and annotated data set of DBT studies that contained 22 032 reconstructed DBT volumes from 5060 patients was made publicly available. A deep learning algorithm for breast cancer detection was developed and tested, with a sensitivity of 65% on a test set.

**Meaning** In this study, the publicly available data set, alongside the deep learning model, could significantly advance the research on machine learning tools in breast cancer screening and medical imaging in general.

+ [Invited Commentary](#)

+ [Supplemental content](#)

Author affiliations and article information are listed at the end of this article.

## Introduction

Deep learning emerged mainly due to rapid increases in access to computational resources and large-scale labeled data.<sup>1</sup> Medical imaging is a natural application of deep learning algorithms.<sup>2</sup> However, well-curated data are scarce, which poses a challenge in training and validating deep learning models. Annotated medical data are limited for a number of reasons. First, the number of available medical images is much lower than the number of available natural images. This is particularly an issue when investigating a condition with fairly low prevalence, such as breast cancer in a screening setting (<1% of screening examinations result in a cancer diagnosis). Second, access to medical imaging data is guided by a number of strict policies given that they contain patients' medical information. Sharing of medical imaging data requires an often nontrivial and time-consuming effort to deidentify the data as well as ensure compliance with requirements from the institution that is sharing the data and beyond. Finally, annotation of medical imaging data typically requires the work of radiologists, who already have high demands on their time.

As a result, the amount of well-annotated large-scale medical imaging data that are publicly available is limited. This is certainly a problem when training deep learning models, but it also results in a lack of transparency when evaluating model performance.

Limited reproducibility of results has been particularly visible in mammography research, arguably the most common radiology application of artificial intelligence (AI) in the last 2 decades.<sup>3-6</sup> Researchers use different, often not publicly available, data sets and solve related but different tasks.<sup>7</sup> Moreover, studies have different evaluation strategies, which makes it difficult to reliably compare methods and results. An AI system must be extensively validated before application in clinical practice. A common shortcoming in many studies is that the test set was obtained from a single institution and a limited number of devices.<sup>8</sup> In addition, some studies make exclusions from the data, which further obscure the true performance of the algorithms.

In this study, we aimed to address some of these challenges. First, we curated and annotated a data set of more than 22 000 three-dimensional (3D) digital breast tomosynthesis (DBT) volumes from 5060 patients. DBT is a new modality for breast cancer screening that, instead of projection images (as in mammography), delivers multiple cross-sectional slices for each breast and offers better performance.<sup>9</sup> We are making this data set publicly available at the Cancer Imaging Archive,<sup>10</sup> a public data hosting service for medical images of various modalities together with community analyses that facilitate the usability of shared data sets. This will allow other groups to improve the training of their algorithms as well as test their algorithms on the same data set, which could improve both the quality of the models and comparison between different algorithms. This could also allow groups that have access to strong machine learning expertise but no clinical data to contribute to the development of clinically useful algorithms.

In addition, we developed and made publicly available a single-phase deep learning model for the detection of abnormal results in DBT that can serve as a baseline for future development or be used for fine-tuning in solving other medical imaging tasks. To our knowledge, this is the first published single-phase deep learning model for DBT. Given that the major challenge of developing the model for this task is a very limited number of positive locations, we evaluated and compared different methods for addressing this issue.

## Methods

### Data Set

This study was approved by the Duke University Health System institutional review board with a waiver of informed consent due its retrospective nature. We analyzed DBT volumes obtained from Duke Health System, following the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline. Specifically, Duke Health Systems Duke Enterprise Data Unified Content Explorer tool was queried to obtain all radiology reports having the word *tomosynthesis* and all

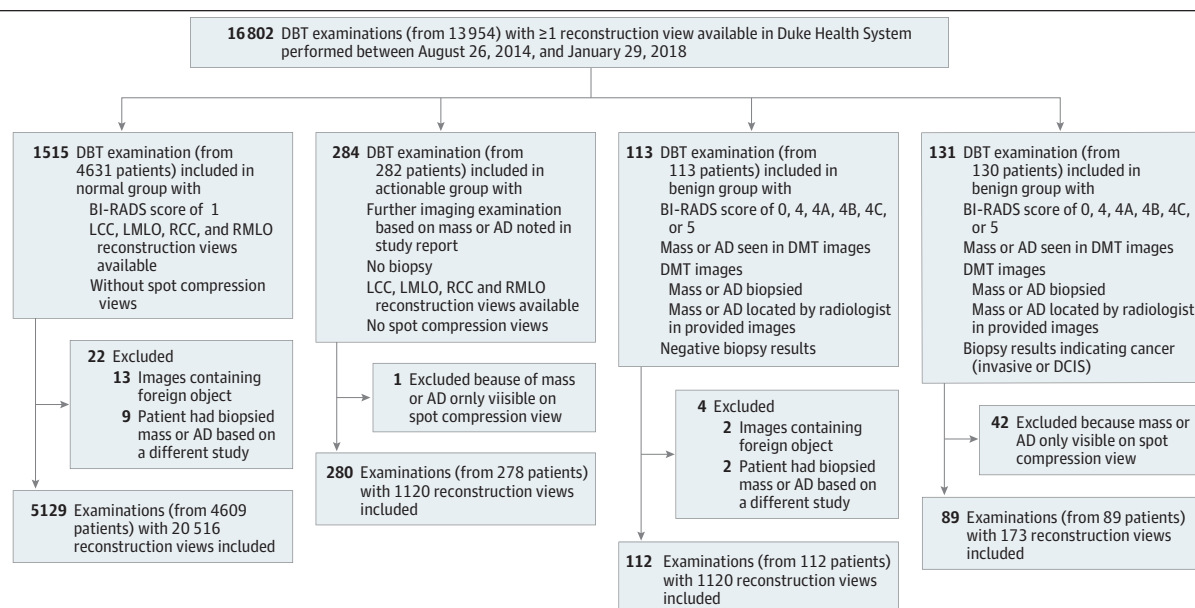
pathology reports having the word *breast* within the search dates of January 1, 2014, to January 30, 2018. The image download based on the study dates and medical record numbers obtained from the radiology reports resulted in an initial collection of 16 802 DBT studies from 13 954 patients performed between August 26, 2014, and January 29, 2018, with at least 1 of the 4 reconstruction volumes (ie, left craniocaudal [LCC], right craniocaudal [RCC], left mediolateral oblique [LMLO], and right mediolateral oblique [RMLO]) available. From this cohort, we divided the studies into 4 groups, as shown in the patient flowchart (Figure 1) and described below.

The normal group included 5129 screening studies from 4609 patients without any abnormal findings that were not subject to further imaging or pathology examinations related to the study in question. Specifically, in this group we included studies that had a Breast Imaging-Reporting and Data System (BI-RADS) score of 1; had LCC, LMLO, RCC, and RMLO reconstruction views available; did not use the words *mass* or *distortion* in the corresponding radiology report, and did not contain spot compression among the 4 views. Spot compression was established based on text processing of radiology reports combined with visual inspection of images. Studies with images containing foreign objects other than implants and markers ( $n = 13$ ) and studies from patients who had biopsied mass or architectural distortion based on a different DBT study ( $n = 9$ ) were excluded.

The actionable group included 280 studies from 278 patients that resulted in further imaging examination based on a mass or architectural distortion noted in the study report. Specifically, we included studies that had a recommendation for a further imaging examination based on a mass or architectural distortion noted in the study report; did not result in a biopsy; had LCC, LMLO, RCC, and RMLO reconstruction views available; and did not contain spot compression among the 4 views. Spot compression was established in the same manner as in the normal group. Studies with images containing foreign objects other than implants and markers ( $n = 2$ ) and studies from patients that had biopsied mass or architectural distortion based on a different DBT study ( $n = 2$ ) were excluded.

The benign group included 112 studies from 112 patients containing benign masses or architectural distortions biopsied based on this DBT examination. Specifically, in this group we included studies that had a BI-RADS score of 0, 4, 4A, 4B, 4C, or 5; had a mass or architectural distortion that was seen in the DBT imaging study in question that was identified using laterality

Figure 1. Patient Flowchart



AD indicates architectural distortion; BI-RADS, Breast Imaging-Reporting and Data System; DBT, digital breast tomosynthesis; LCC, left craniocaudal; LMLO, left mediolateral oblique; RCC, right craniocaudal; RMLO, right mediolateral oblique.

and/or location noted in a related breast pathology report and was biopsied; had benign results of all biopsies per the pathology reports; and a radiologist was able to retrospectively locate at least 1 of the biopsied benign masses or architectural distortions in the reconstruction views from the study. One study for which the biopsied mass was visible only on spot compression views was excluded.

The cancer group included 89 studies from 89 patients with at least 1 cancerous mass or architectural distortion that was biopsied based on this DBT examination. Specifically, we included studies that had a mass or architectural distortion seen in the DBT images that was identified using laterality and/or location noted in a related breast pathology report and was biopsied; had at least 1 biopsied mass or architectural distortion corresponding to cancer (invasive or ductal carcinoma in situ) per the pathology report; and a radiologist was able to retrospectively locate at least 1 of the biopsied cancerous mass or architectural distortion in the reconstruction views from the study. Studies for which all cancerous masses or architectural distortions were visible only on spot compression views ( $n = 42$ ) were excluded. More details on the exclusion of cases from the initial population are provided in eAppendix 1 in the [Supplement](#).

### Training, Validation, and Test Sets

In total, our data set contained 22 032 reconstructed volumes that belonged to 5610 studies from 5060 patients. It was randomly split into training, validation, and test sets in a way that ensured no overlap of patients between the subsets. The test set included 460 studies from 418 patients. For the validation set, we selected 312 studies from 280 patients, and the remaining 4838 studies from 4362 patients were in the training set. The selection of cases from the benign and cancer groups into the test and validation sets was performed to assure a similar proportion of masses and architectural distortions. Descriptive statistics for all the subsets are provided in **Table 1**.

### Image Annotation

Study images along with the corresponding radiology and pathology reports for each biopsied case were shown to 2 radiologists at our institution (R.W. and S.G.) for annotation. We asked the radiologists to identify masses and architectural distortions that were biopsied and to put a rectangular box enclosing them in the central slice using a custom software developed by a researcher (N.L.) in our laboratory. Each case was annotated by 1 of 2 experienced radiologists. The first radiologist, with 25 years of experience in breast imaging (R.W.), annotated 124 cases, whereas the second radiologist, with 18 years of experience in breast imaging (S.G.), annotated 77 cases. This way we obtained 190 bounding boxes for cancerous lesions in 173 reconstruction views and 245 bounding boxes for benign lesions in 223 reconstruction views. There were 336 and 99 bounding boxes for masses and architectural distortions, respectively, across cancerous and benign lesions.

**Table 1. Descriptive Statistics of the Data Set Used for Training, Validation, and Testing**

Characteristics	No.		
	Training set	Validation set	Test set
Patients			
Total	4362	280	418
Normal group, No. (%)	4109 (94.2)	200 (71.4)	300 (71.8)
Actionable group, No. (%)	178 (4.1)	40 (14.2)	60 (18.9)
Benign group, No. (%)	62 (1.4)	20 (7.1)	30 (7.2)
Cancer group, No. (%)	39 (0.9)	20 (7.1)	30 (7.2)
Studies	4838	312	460
Reconstruction volumes	19 148	1163	1721
Bounding boxes for cancerous lesions	87	37	66
Bounding boxes for benign lesions	137	38	70
Bounding box diagonal, mean (SD), pixels	344 (195)	307 (157)	317 (166)

## Baseline Algorithm

### Preprocessing

First, we applied a basic preprocessing by window leveling images based on information from the Digital Imaging and Communications in Medicine file header. Then, each slice was downsampled by a factor of 2 using  $2 \times 2$  local mean filter to reduce computational and memory footprint. After that, we eroded nonzero image pixels with a filter of 5-pixel radius for skin removal. Finally, we extracted the largest connected component of nonzero pixels for segmenting the breast region.

### Detection Algorithm

For a baseline method to detect lesions, we used a single-phase fully convolutional neural network for 2-D object detection<sup>11</sup> with DenseNet<sup>12</sup> architecture. The model processes each 2-D input slice independently. Following this,<sup>11</sup> raw model predictions correspond to a grid in the input slice image with cells sized  $96 \times 96$  pixels. For each cell, the network outputs a confidence score for containing the center point of a box and 4 values defining the location and dimensions of the predicted box. A bounding box is defined by offset from the cell center point as well as scale in relation to a square anchor box sized  $256 \times 256$  pixels.<sup>13</sup> Each cell was restricted to predicting exactly 1 bounding box.

The network was optimized using Adam,<sup>14</sup> with an initial learning rate of 0.001 and batch size of 16 for 100 epochs over positive examples and early stopping strategy with a patience of 25 epochs. Weights were randomly initialized using the Kaiming method,<sup>15</sup> and biases in the last layer were set according to Lin et al.<sup>16</sup>

For training, we sampled positive slices containing ground truth boxes from volumes belonging to the biopsied groups. The number of positive slices (ie, slices containing a tumor) was established as the square root of the average dimension in pixels of the box drawn by a radiologist on the center slice of the tumor. The ground truth 3-D box was defined by the 2-D rectangle drawn by the radiologist with the third dimension defined by the number of slices, as described previously. Then, we randomly cropped a slice image to a size of  $1056 \times 672$  pixels, which resulted in an output grid sized  $11 \times 7$  pixels so that the cropped slice image included the entire ground truth bounding box. For validation, the slice span of ground truth boxes was reduced by a factor of 2 compared with the training phase, and we fixed selected slice and cropped slice image regions for each case. This was done to ensure comparable validation performance was measured based on the same input slice for all runs and across epochs. All hyperparameters and algorithmic strategies described previously were decided on the validation set.

During inference, we used entire image slices as the input and padded them with zeros when necessary to match the label grid size. To obtain predictions for a volume, we split it into halves and combined slice-based predictions for each half by averaging them. Then, we applied the following postprocessing. First, predicted boxes for which fewer than half the pixels were in the breast region were discarded to eliminate false-positive predictions outside of the breast. Then, we applied a nonmaximum suppression algorithm<sup>17</sup> by merging all pairs of predicted boxes that had a confidence score ratio of less than 10 and an intersection over union greater than 50%. The confidence score of a resulting box was a maximum of scores from the 2 merged boxes.

## Experiments

To provide an insight into the effects of different hyperparameters on the performance, we performed a grid search over different network sizes and objectness loss functions that address the problem of class imbalance.<sup>18</sup> Our problem was characterized by a significant imbalance between the bounding boxes corresponding to lesions and background class that the network learns to distinguish in the training process. The 4 tested loss functions for addressing this problem were: (1) binary cross-entropy, (2) weighted binary cross-entropy, (3) focal loss,<sup>16</sup> and (4) reduced focal loss.<sup>19</sup> Weighted binary cross-entropy assigns different weights to positive and negative examples based on class prevalence. Focal loss is a parametrized loss function that reduces the importance of examples that are correctly classified without high confidence, as shown in eAppendix 1 in the [Supplement](#).

Finally, reduced focal loss is equivalent to binary cross-entropy for examples misclassified with a confidence lower than 0.5, and after this threshold, loss value is gradually reduced to focal loss. For bounding box localization loss, we used mean squared error as in Redmon et al.<sup>11</sup> In total, we trained 768 models, and the results from all runs are provided in eAppendix 2 in the Supplement. The code for all experiments and network architecture together with the trained model weights are publicly available.<sup>20</sup>

In the grid search, model selection was based on the sensitivity at 2 false positives per slice computed on the validation set after every epoch. For each loss function, we selected the best performing model for 3-D evaluation on the entire validation set. Following this 3-D evaluation, the model with the highest sensitivity at 2 false positives per DBT volume on the validation set was used to generate predictions on the test set for the final evaluation.

Final Model Evaluation on the Test Set

For the final evaluation of the baseline detection algorithm, we used the free-response receiver operating characteristic (FROC) curve, which shows the sensitivity of the model in relation to the number of false-positive predictions placed in slice images, volumes, or cases. A predicted box was considered a true positive if the distance between its center point and the center of a ground truth box was either smaller than half of the ground truth box diagonal or smaller than 100 pixels. The additional 100 pixels condition was implemented to prevent punishing correct detections for very small lesions with unclear boundaries. In terms of the third dimension, the ground truth bounding box was assumed to span 25% of volume slices before and after the ground truth center slice, and the predicted box center slice was required to be included in this range to be considered a true positive.

In addition to the volume-based evaluation described above, we evaluated the accuracy of model predictions using breast-based FROC. In this case, a prediction for a breast was considered true positive if any lesion on any view for this breast was detected according to the criteria described above. This metric most accurately reflects the model performance in a clinical setting.

Statistical Analysis

For the final evaluation of the baseline detection algorithm, we used the FROC curve, which shows the sensitivity of the model in relation to the number of false-positive predictions placed in slice images, volumes, or cases. Sensitivity values are reported together with 95% CIs, which were computed using bootstrapping with 2000 bootstraps. For this, we used an open-source statistical tool implemented in Python.<sup>21</sup>

Results

The number of patients in the data set was 5060, with 5059 women (100.0%) and 1 man (<0.1%). The mean (SD) age at the date of patient's first examination included in our data set was 55 (11) years. Age statistics were computed based on 5059 patients. The date of birth for 1 patient was unknown. Table 2 provides demographic characteristics for patients in our data set.

Performance on the Validation Set

All tested loss functions performed similarly, with the best configuration for each loss achieving greater than 78% sensitivity at 2 false positives per slice. Using the best model from the grid search for each loss function in the 2-D per-slice evaluation, we ran inference and evaluated selected models on the entire validation set using the 3-D per-volume evaluation. The best performance, with 60% sensitivity at 2 false positives per DBT volume, was achieved by the network trained using focal loss. In comparison, sensitivity at the same threshold achieved by binary cross-entropy and weighted binary cross-entropy was 59%, whereas reduced focal loss obtained 58%. The model trained using

Table 2. Characteristics of Patients in the Data set

Characteristic	Participants, No. (%)
Age, mean (SD), y	55 (11)
Missing age	1 (<0.1)
Sex	
Women	5059 (100.0)
Men	1 (<0.1)
Race	
White	3700 (73.1)
Black or African American	957 (18.9)
Asian	180 (3.6)
American Indian or Alaskan Native	11 (0.2)
Native Hawaiian or other Pacific Islander	2 (<0.1)
Other <sup>a</sup>	52 (1.0)
≥2 races	56 (1.1)
Not reported, declined, or unavailable	102 (2.0)

<sup>a</sup> The other category was present in the original data, and it was not specified what groups were included.

focal loss was selected for evaluation on the test set. More details on the grid search results and FROC curves on the validation set are provided in eAppendix 1 in the [Supplement](#).

### Performance on the Test Set

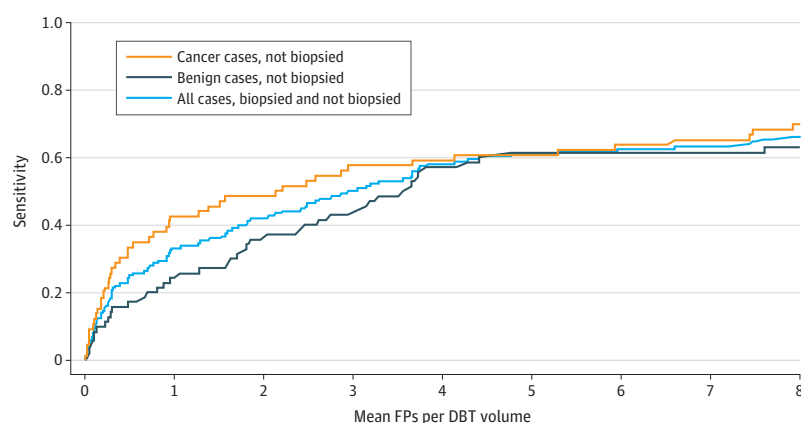
Using a model trained by optimizing focal loss function, we generated predictions for the test set. The model achieved a sensitivity of 42% (95% CI, 35%-50%) at 2 false positives per DBT volume as shown on the FROC curve in **Figure 2**. Better performance was reached on the cancer cases than on benign cases.

Finally, we evaluated the selected model using breast-based FROC computed on the test set. In this case, sensitivity at 2 false positives per DBT volume for test cases with cancer and all test cases was 67% (95% CI, 53%-80%) and 65% (95% CI, 56%-74%), respectively. The breast-based FROC curve for the test set is shown in **Figure 3**.

## Discussion

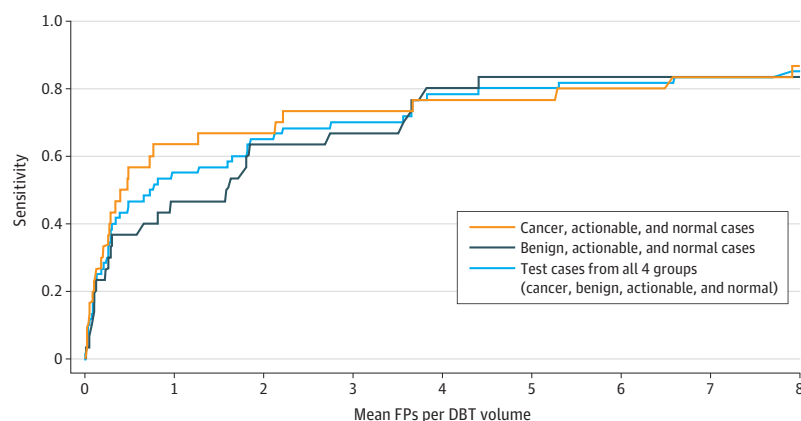
In this study, we described a large-scale data set of DBT examinations containing data for 5060 patients that we shared publicly. We also trained the first single-phase detection model for this data set that will serve as a baseline for future development.

**Figure 2. Free-Response Receiver Operating Characteristic Curve Showing Performance on the Test Set of a Model Trained Using Focal Loss**



DBT indicates digital breast tomosynthesis; FP, false positive.

**Figure 3. Breast-Based Free-Response Receiver Operating Characteristic Curve for the Test Set**



DBT indicates digital breast tomosynthesis; FP, false positive.



Our study included annotations for both masses and architectural distortions. Those abnormal findings appear differently in DBT images and therefore constitute a more challenging task for an automated algorithm. A model that focuses on a single task (such as many previously published models for breast imaging) could show overoptimistic performance. This more inclusive data set more accurately represents true clinical practice of breast cancer screening. Furthermore, our data set, which includes normal and actionable cases, is representative of a screening cohort.

Our detection model was developed using only 124 and 175 bounding boxes for cancerous and benign lesions, respectively. No pretraining on other data sets or similar modalities was used. In addition, our detection method is a single-phase deep convolutional neural network, which does not require multiple steps for generating predictions. We showed that a moderate performance can be achieved with a limited training data. In comparison, a previous study<sup>22</sup> reported sensitivity less than 20% at 2 false positives per volume for a model trained from scratch using only DBT data without pretraining on a much larger data set of mammograms. In another study,<sup>23</sup> a sensitivity of greater than 80% at 2 false positives per volume was reached for a data set containing only architectural distortions. In Fan et al,<sup>24</sup> a 3-D deep learning model was developed that achieved 90% sensitivity at 0.8 false positives per volume on a data set containing only abnormal images with masses.

The methods for evaluating performance of detection algorithms vary. The method used in this study is robust to models predicting large bounding boxes as opposed to evaluation methods that consider a predicted box as a true positive if it contains the center point of the ground truth box. In our study, the center point of the predicted box was required to be contained in the ground truth box as well. Furthermore, we were solving a 3-D detection task, which generates a higher number of false positives than 2-D detection tasks. While the performance of our model is not comparable with the performance of radiologists, our goal was to set a baseline for a model that is trained only on the provided data and without access to large-scale computer clusters.

## Limitations

This study had limitations. First, the data set contains images that were collected from a single institution. Second, we did not include annotations for calcifications and/or microcalcifications because they are notably different visual structures in the context of a computer vision detection system. Detection of calcifications was outside of our research goals when assembling this data set. This may produce a different composition of DBT volumes than typically encountered in a clinical setting. Third, the number of biopsied cases was much smaller than the number of images without bounding boxes. However, this reflects the prevalence of cancers in screening populations.

Images in the data set were interpreted by several radiologists, and the assignment of studies to groups was made, among other criteria, based on BI-RADS score, which is known to have high interreader variability. Moreover, for the first 6 to 12 months of DBT adoption at our institution, radiologists relied on both DBT and mammography for BI-RADS score assignment, and they gradually moved to diagnosis based on DBT and C-view.

Given that our criteria for a normal examination was the assessment of a radiologist for that examination, there exists a slight possibility that a cancer was detected in a follow-up examination that could then be retrospectively visible on the examination that was considered normal. However, this is a highly unlikely scenario.

Additionally, our baseline model achieved slightly better performance on test cases from the cancer group compared with the benign group. This could be explained by the fact that cancerous lesions in our data set or in general are easier to detect by a computer vision algorithm.

## Conclusions

In this study, we curated and annotated a publicly available data set of DBT volumes for future training and validation of AI tools. All the factors described previously make this data set a challenging but realistic benchmark for the future development of methods for detecting masses



and architectural distortions in DBT volumes. These factors, including different types of abnormal results, exclusions of different types of cases, and different evaluation metrics, make it difficult to compare our method with those previously presented in the literature.<sup>22,25,26</sup> This further underlines the importance of the data set shared in this study.

## ARTICLE INFORMATION

**Accepted for Publication:** May 27, 2021.

**Published:** August 16, 2021. doi:[10.1001/jamanetworkopen.2021.19100](https://doi.org/10.1001/jamanetworkopen.2021.19100)

**Open Access:** This is an open access article distributed under the terms of the [CC-BY License](#). © 2021 Buda M et al. *JAMA Network Open*.

**Corresponding Author:** Mateusz Buda, MSc, Department of Radiology, Duke University Medical Center, 2424 Erwin Rd, Durham, NC 27710 ([buda@kth.se](mailto:buda@kth.se)).

**Author Affiliations:** Department of Radiology, Duke University Medical Center, Durham, North Carolina (Buda, Saha, Walsh, Ghate, Li, Świąćicki, Lo, Mazurowski); Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina (Lo, Mazurowski); Department of Computer Science, Duke University, Durham, North Carolina (Mazurowski); Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina (Mazurowski).

**Author Contributions:** Mr Buda and Dr Mazurowski had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

**Concept and design:** Buda, Saha, Li, Lo, Mazurowski.

**Acquisition, analysis, or interpretation of data:** Buda, Saha, Walsh, Ghate, Li, Świąćicki, Mazurowski.

**Drafting of the manuscript:** Buda, Saha, Li.

**Critical revision of the manuscript for important intellectual content:** All authors.

**Statistical analysis:** Buda, Saha, Li.

**Obtained funding:** Mazurowski.

**Administrative, technical, or material support:** Buda, Saha, Walsh, Ghate, Świąćicki, Lo, Mazurowski.

**Supervision:** Saha, Li, Świąćicki, Mazurowski.

**Image interpretation and documentation:** Ghate.

**Conflict of Interest Disclosures:** Dr Walsh reported receiving personal fees from Therapixel outside the submitted work. Dr Ghate reported receiving personal fees from Therapixel during the conduct of the study and personal fees from Siemens outside the submitted work. Dr Mazurowski reported serving as an advisor to Gradient Health outside the submitted work. No other disclosures were reported.

**Funding/Support:** This work was supported by a grant 1 R01 EBO21360 from the National Institutes of Health to Dr Mazurowski.

**Role of the Funder/Sponsor:** The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

## REFERENCES

1. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 25*. Accessed July 7, 2021. <https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>
2. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017; 42:60-88. doi:[10.1016/j.media.2017.07.005](https://doi.org/10.1016/j.media.2017.07.005)
3. Le EPV, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. *Clin Radiol*. 2019;74(5): 357-366. doi:[10.1016/j.crad.2019.02.006](https://doi.org/10.1016/j.crad.2019.02.006)
4. Schaffter T, Buist DSM, Lee CI, et al; and the DM DREAM Consortium. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open*. 2020;3(3): e200265-e200265. doi:[10.1001/jamanetworkopen.2020.0265](https://doi.org/10.1001/jamanetworkopen.2020.0265)
5. Kim H-E, Kim HH, Han B-K, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health*. 2020;2(3):e138-e148. doi:[10.1016/S2589-7500\(20\)30003-0](https://doi.org/10.1016/S2589-7500(20)30003-0)

6. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6
7. Geras KJ, Mann RM, Moy L. Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. *Radiology*. 2019;293(2):246-259. doi:10.1148/radiol.2019182627
8. AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing. *Med Phys*. 2018;45(3):1150-1158. doi:10.1002/mp.12752
9. Vedantham S, Karellas A, Vijayaraghavan GR, Kopans DB. Digital breast tomosynthesis: state of the art. *Radiology*. 2015;277(3):663-684. doi:10.1148/radiol.2015141303
10. Breast cancer screening—digital breast tomosynthesis (BCS-DBT). Cancer Imaging Archive. June 9, 2021. Accessed July 14, 2021. <https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=64685580>
11. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: 2016 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016:779-788. doi:10.1109/CVPR.2016.91
12. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2017:2261-2269. doi:10.1109/CVPR.2017.243
13. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. Accessed July 7, 2021. <https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf>
14. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint updated January 30, 2017. Accessed July 7, 2021. <https://arxiv.org/abs/1412.6980>
15. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE; 2015:1026-1034. doi:10.1109/ICCV.2015.123
16. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. arXiv. Preprint updated February 7, 2018. Accessed July 7, 2021. <https://arxiv.org/abs/1708.02002>
17. Neubeck A, Van Gool L. Efficient non-maximum suppression. In: 18th International Conference on Pattern Recognition (ICPR'06). IEEE; 2006:850-855. doi:10.1109/ICPR.2006.479
18. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw*. 2018;106:249-259. doi:10.1016/j.neunet.2018.07.011
19. Sergievskiy N, Ponamarev A. Reduced focal loss: 1st place solution to xView object detection in satellite imagery. arXiv. Preprint updated April 25, 2019. Accessed July 7, 2021. <https://arxiv.org/abs/1903.01347>
20. Duke DBT data. Github. Accessed July 14, 2021. <https://github.com/MaciejMazurowski/duke-dbt-data>
21. Machine learning statistical utils. Github. Accessed July 14, 2021. <https://github.com/mateuszbudam/ml-stat-util>
22. Samala RK, Chan H-P, Hadjiiski L, Helvie MA, Wei J, Cha K. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. *Med Phys*. 2016;43(12):6654-6666. doi:10.1118/1.4967345
23. Li Y, He Z, Lu Y, et al. Deep learning of mammary gland distribution for architectural distortion detection in digital breast tomosynthesis. *Phys Med Biol*. 2021;66(3):035028. doi:10.1088/1361-6560/ab98d0
24. Fan M, Zheng H, Zheng S, et al. Mass detection and segmentation in digital breast tomosynthesis using 3D-mask region-based convolutional neural network: a comparative analysis. *Front Mol Biosci*. 2020;7:599333. doi:10.3389/fmolb.2020.599333
25. Mendel K, Li H, Sheth D, Giger M. Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography. *Acad Radiol*. 2019;26(6):735-743. doi:10.1016/j.acra.2018.06.019
26. Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med*. 2021;27(2):244-249. doi:10.1038/s41591-020-01174-9

#### SUPPLEMENT.

eAppendix 1. Supplemental Methods

eReferences.

eAppendix 2. Results From All Model Runs

## **A.9 A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis**



OPEN

# A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis

Albert Swiecicki<sup>1</sup>✉, Nicholas Konz<sup>1</sup>, Mateusz Buda<sup>2</sup> & Maciej A. Mazurowski<sup>1,2</sup>

Deep learning has shown tremendous potential in the task of object detection in images. However, a common challenge with this task is when only a limited number of images containing the object of interest are available. This is a particular issue in cancer screening, such as digital breast tomosynthesis (DBT), where less than 1% of cases contain cancer. In this study, we propose a method to train an inpainting generative adversarial network to be used for cancer detection using only images that do not contain cancer. During inference, we removed a part of the image and used the network to complete the removed part. A significant error in completing an image part was considered an indication that such location is unexpected and thus abnormal. A large dataset of DBT images used in this study was collected at Duke University. It consisted of 19,230 reconstructed volumes from 4348 patients. Cancerous masses and architectural distortions were marked with bounding boxes by radiologists. Our experiments showed that the locations containing cancer were associated with a notably higher completion error than the non-cancer locations (mean error ratio of 2.77). All data used in this study has been made publicly available by the authors.

Deep learning methods have been shown to be highly successful in the analysis of medical images<sup>1</sup>. However, typically a large amount of data is needed to train accurate models. The collection of a large numbers of cases is particularly challenging when attempting to work with rare diseases. In screening populations, the prevalence of some diseases can be as low as 1%, resulting in a large number of normal exams, yet very few exams depicting abnormalities. One of the domains where we can observe such low prevalence is mammography, imaging exams intended to detect breast cancer in otherwise healthy women. Based on<sup>2</sup>, only 9812 out of 1,682,504 screening mammograms examinations performed between 2007 and 2013 consisted of cancerous alternations, resulting in an approximately 0.6% ratio between positive and negative test results. The three-dimensional, more modern form of mammography, called digital breast tomosynthesis (DBT) may find a slightly larger number of cancers since it provides better lesion visibility when compared with analog mammography or full-field digital mammography (FFDM)<sup>3</sup>. However, the prevalence of abnormal results remains very low.

Such imbalance in the training dataset causes significant problems when training deep learning algorithms and has been shown to negatively affect model performance<sup>4</sup>. In detection tasks, training difficulty already arises from the very limited number of images that contain abnormalities, but as in the case of mammography, this is made even worse when combined with the fact that the abnormalities themselves occupy relatively small parts of the images. Therefore, in order to make some sort of meaningful training progress, it becomes crucial to effectively utilize images that do *not* contain abnormalities, which *are* available in abundance.

Current supervised deep learning-based detection algorithms are not well-designed to take advantage of images that do not contain abnormalities. Images without abnormalities are used in anomaly detection algorithms where models try to learn data distributions and, based on normal data, try to predict unusual behaviors. One of the approaches to utilizing images with no abnormalities is to extract feature representation from normal data before training models with rare abnormal data. The most popular ways of extracting features generally (1) use

<sup>1</sup>Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. <sup>2</sup>Department of Radiology, Duke University, Durham, NC, USA. ✉email: albert.swiecicki@duke.edu

Set	Type	Patients	Studies	Volumes	Slices/images
Train	Normal	4109	4558	18,232	91,160
Validation	Normal	200	232	928	4640
Test	Cancer	39	39	70	70

**Table 1.** Data used during experiments.

a compression-decompression network called an *autoencoder*<sup>5,6</sup> or (2) involve generative adversarial networks (GANs) to learn data distributions<sup>7,8</sup>.

Our hypothesis is that breasts, similarly to many other objects, have a certain expected structure visible within images. Radiologists learn this structure by viewing thousands of breast images. Once structure is learned, an abnormality can be detected as a location where the tissue looks different than expected. Following this hypothesis, we propose to simulate this phenomenon using a computer algorithm. Specifically, we developed an algorithm that is able to fill in a missing part of an image, at a given location, with what is expected based on the rest of the image and based on what the algorithm has seen in tens of thousands of other images that don't contain abnormalities. A state-of-the-art generative adversarial network (GAN) is used for this image completion task. A recent study<sup>9</sup> have shown that image completion algorithms are able to complete images with high-quality patches consistent with their surroundings<sup>9</sup>. Then, if the expected image at this location is different from the actual image, the location is considered suspicious.

The purpose of this research is to determine whether the model trained on data without abnormalities will have difficulty with reconstructing previously unseen abnormal structures. The hypothesis is validated on a set of 70 digital breast tomosynthesis images containing cancerous lesions, by measuring completion error inside and outside of bounding boxes and visualizing model losses in the form of heatmaps.

While GANs have been previously used in the context of anomaly detection<sup>10</sup>, we are familiar with only one study that uses neural network-based image completion for this purpose. Specifically, in a study conducted by Haselmann et al.<sup>11</sup>, mean-squared error (MSE) was incorporated with a GAN to perform image completion (inpainting) abnormality detection, showing promising results but only on a relatively easy task. In this study, the difference between the original image and the completed image (measured by MSE) was used to determine whether a particular location is likely to be abnormal. Here, we extend this study by applying the concept into much more challenging space of medical imaging, introducing a newer attention model for image completion<sup>9</sup> and evaluating the performance of the model using different mask sizes, model input sizes, and losses on non-trivial medical data.

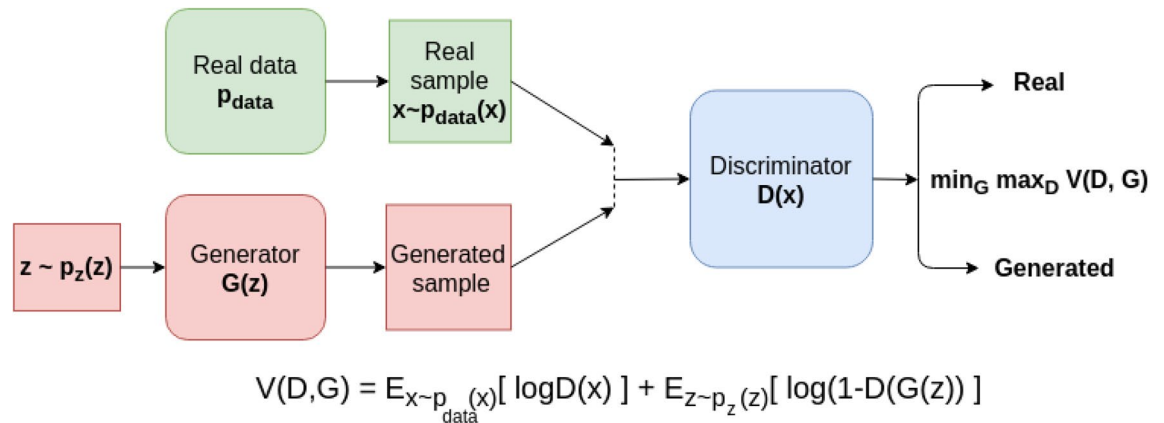
This study has multiple contributions in terms of its technical aspect and the application. It is the first study that attempts to use image completion for abnormality detection in the context of medical imaging. This comes with a variety of challenges including high resolution images (approximately 50 times more pixels in an image than in<sup>11</sup>). We also introduce the generative image inpainting with contextual attention model<sup>9</sup> in the context of anomaly detection. Additionally, we use the discrimination loss measure to determine abnormal-looking locations in the context of image completion. Finally, we explore the impact of hyperparameters, such as the field of view and mask size, on the performance of the algorithm.

## Methods

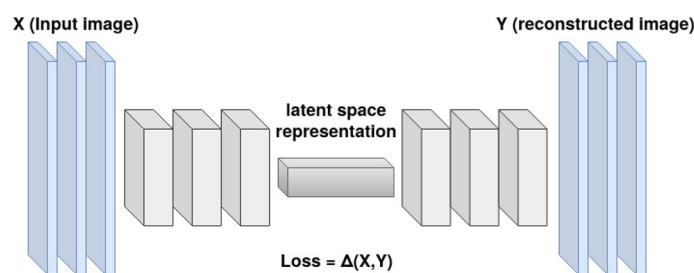
**Dataset.** In this study, we used a dataset of digital breast tomosynthesis (DBT) screening studies gathered from the Duke Health System. It contained 4829 studies collected from 4348 patients resulting in 19,230 reconstruction volumes. There are two types of cases within the study: (1) normal and (2) cancer. For the cancer group, lesion bounding boxes were provided by radiologists from Duke Hospital. In the normal group, every study consists of left and right cranial-caudal (CC) and mediolateral-oblique (MLO) views. Studies in the cancer group consist of one or more CC or/and MLO views. Studies with spot compression were not included in our dataset.

The normal set was randomly divided (by patient) into two exclusive training and validation sets with 18,232 and 928 reconstruction views respectively. In addition, we used 70 volumes from the cancer group to evaluate our algorithm in the context of abnormality detection. Six cases where the abnormality was contained within a small distance (128 pixels) from the edge of the image were removed to arrive at the 70 used volumes. From each volume in the normal set (training and validation), we took five random slices/images. From cancer set volumes we only used the slice where radiologist placed a bounding box; if more than one abnormality was marked (which occurred in eight volumes), we selected one slice randomly from the subset of marked slices. The number of cases used for training, validation and testing are shown in Table 1. All data used in this study will be made publicly available on The Cancer Imaging Archive. The retrospective clinical data collection was approved by the Institutional Review Board (IRB) of the Duke University Health System (DUHS), and the methods used in this study were carried out in accordance with relevant guidelines and regulations. The requirement for informed consent was waived by the DUHS IRB.

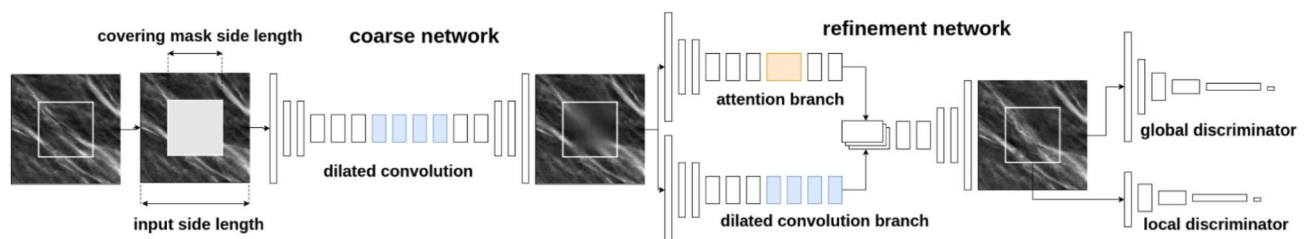
**Generative adversarial networks.** Generative adversarial networks (GANs) introduced in<sup>7</sup>, are based on the idea of two networks competing with each other. One of the networks is responsible for the generation of “fake” training data that appears to be real, by learning to approximate the distribution that generated the real training data. This network is called a generator, denoted  $G(z)$ , because it takes a vector of random noise  $z$  as input, and maps it to a generated datapoint (image, in our case). The second network, called the *discriminator*, or



**Figure 1.** A standard GAN architecture and loss function.



**Figure 2.** Autoencoder architecture.



**Figure 3.** Model architecture and image completion process diagram.

*critic*, is used to distinguish between generated and true samples. It is labeled  $D(x)$ , as the network takes a sample  $x$  and outputs the probability of  $x$  being from the real dataset. The competition between the two networks can be described as a min–max game of two players trying to beat each other, described in Fig. 1.

GANs with multiple convolutional layers are called deep convolutional GANs (DCGANs)<sup>8</sup> and are used amongst other methods for the generation of realistic images<sup>12</sup>, image denoising<sup>13</sup>, image translation<sup>14</sup> and image completion<sup>9</sup>. Image completion is often performed using generators of architecture similar to autoencoders<sup>15</sup>, which foster learning a latent representation of the data. Latent data representation is achieved by compressing and decompressing input data in the way which minimizes information decline. Figure 2 demonstrates sample autoencoder architecture.

**Image completion task and the architecture.** In the task of image completion, a part of an image is covered and the model attempts to reconstruct it based on the parts of the image that are present. We assume that the missing part of the image, the *mask*, is square, and we refer to the size of the missing part as the *mask size*. We approached the task of image completion using DCGAN architecture with a two-phase generator followed by local and global discriminators<sup>9</sup>. In order to train the image completion model, we cover part of the image and recreate the covered part using the generator, based on the remainder of the image. We then use the discriminator to estimate the probability of the generated patch being real.

The architecture of the model and a diagram of the image completion process are shown in Fig. 3. In the first stage of the generator, a *coarse* network constructed from dilated convolutional blocks is used to create an imperfect, blurred prediction for the missing patch of the image. The second part of the generator, the *refinement* network, improves the quality of the completed region with more fine-grained details using combined *contextual*



*attention* and dilated convolutional branches. The contextual attention branch, created by<sup>9</sup>, optimizes consistency between the inferred missing patch and the rest of the surrounding image, hereafter the *field of view*, by examining the inner product/cosine-similarity of features within the generated missing patch and features found in the surroundings. Local and global discriminators are responsible for achieving consistency between the completed masked region and the entire image. We experimented with the following parameters: (i) a mask size of  $64 \times 64$  and  $128 \times 128$  pixels, and (ii), a field of view of  $256 \times 256$  and  $512 \times 512$  pixels. To obtain the same dimensionality of feature representation for both of the tested field of view sizes, we append a convolutional layer to the input of the global discriminator module when the field of view size is  $512 \times 512$  pixels.

**Training details.** While the min-max objective that dictates the training of GANs (Fig. 1) is conceptually simple, in practice training GANs to give usable results is a difficult task. The goal of generative modeling is essentially to make the “fake” data distribution that the generator learns to sample from,  $P_g$ , as similar as possible to the real data distribution  $P_r$ . However, if one tries to do this using common distribution divergence/distance metrics, such as the Kullback–Leibler (KL) divergence, that are usually used to train GANs, this optimization procedure is often practically difficult, due to issues such as discontinuities and/or vanishing gradients within the objective function with respect to the network’s parameters, that can occur when a real sample is not within the support of  $P_r$ . The Wasserstein distance, described shortly, was proposed as a solution to these problems<sup>16</sup>, and is a key component of our model’s loss function.

The Wasserstein distance between two distributions can intuitively be thought of as the minimal effort needed to transport probability mass between these distributions; it is theoretically defined as

$$W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} E_{(x,y) \sim \gamma} [\|x - y\|], \quad (1)$$

where  $\Pi(P_r, P_g)$  is the set of all distributions  $\gamma$  whose marginal distributions are  $P_r$  and  $P_g$ . This equation is unsurprisingly practically intractable, but a more useful form of it can be obtained using the *Kantorovich-Rubinstein duality*<sup>16</sup>, which gives

$$W(P_r, P_g) = \sup_{f \in \mathcal{F}} E_{x \sim P_r} [f(x)] - E_{\bar{x} \sim P_g} [f(\bar{x})], \quad (2)$$

where  $\mathcal{F}$  is the set of all 1-Lipschitz functions. Practically speaking, using  $W(P_r, P_g)$  as the distance measure for training a GAN will modify the min-max objective function (Fig. 1) to become

$$\min_G \max_{D \in \mathcal{F}} E_{x \sim P_r} [D(x)] - E_{\bar{x} \sim P_g} [D(\bar{x})]. \quad (3)$$

An important note here is that the discriminator  $D$  is constrained to be 1-Lipschitz, which can be thought of as forcing the high-dimensional analog of the “slope” of  $D$  with respect to its network parameters to be no greater than 1. Arjovsky et al.<sup>16</sup> originally implemented this constraint by “clipping” the weights of  $D$  to be within a certain magnitude, but this can lead to undesirable training instability. As such, we utilize WGAN-GP in our model, an improved version of the WGAN introduced by<sup>17</sup> that instead enforces the 1-Lipschitz constraint by adding a *gradient penalty* term

$$\lambda E_{\hat{x} \sim P_{\hat{x}}} (\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2 \quad (4)$$

to the objective function, where  $\hat{x}$  are sampled from the straight line between points sampled from  $P_r$  and  $P_g$  and  $\lambda$  is a constant hyperparameter. Essentially what this added term does is instead enforce the aforementioned constraint by penalizing the size of the gradient of  $D$  with respect to its input, which gives improved training performance.

We can now write the total loss function  $\mathcal{L}_{\text{total}}$  for our model as the sum of individual loss components, as

$$\mathcal{L}_{\text{total}} = \alpha_{\text{mask}} \mathcal{L}_{\text{mask}} + \alpha_{\text{FOV}} \mathcal{L}_{\text{FOV}} + \alpha_{\text{GAN}} \mathcal{L}_{\text{WGAN},G} + \mathcal{L}_{\text{WGAN},D} + \lambda \mathcal{L}_{\text{WGAN-GP}}, \quad (5)$$

where:

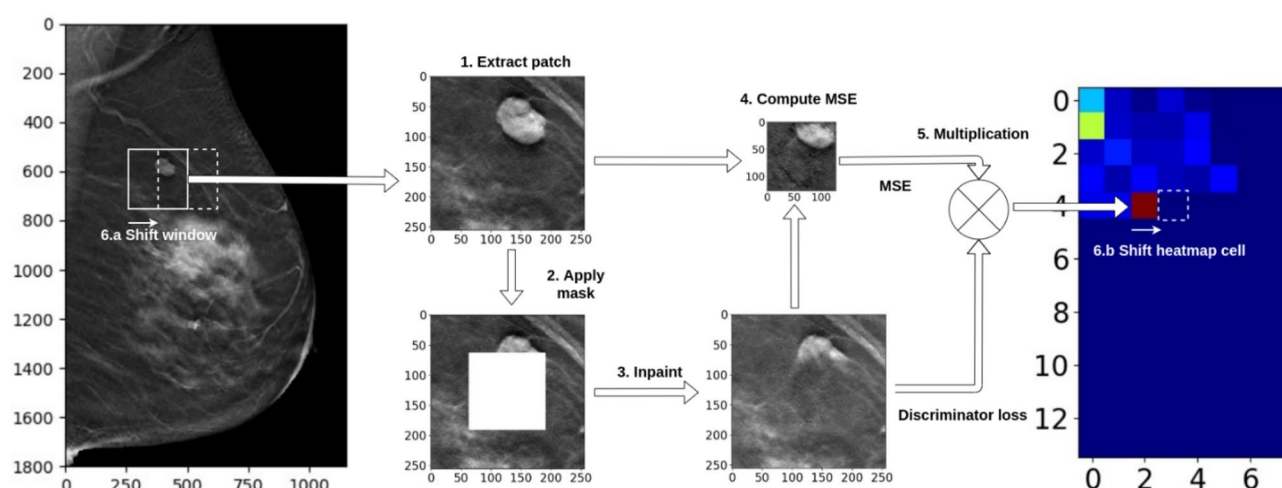
$\mathcal{L}_{\text{mask}}$  is the  $L^1$  (Manhattan) distance between the coarse prediction for the masked region and the corresponding region of the ground truth, added to the same for the fine prediction,

$\mathcal{L}_{\text{FOV}}$  is the  $L^1$  distance between the coarse prediction for the non-masked part of the image and the corresponding ground truth, added to the same for the fine prediction,

$\mathcal{L}_{\text{WGAN},G}$  and  $\mathcal{L}_{\text{WGAN},D}$  are the WGAN losses between the local and global discriminators and the two-stage generator (see Eq. 3), with  $\mathcal{L}_{\text{WGAN-GP}}$  being the added gradient penalty (GP) terms for both discriminators (see Eq. 4).

Finally,  $\alpha_{\text{mask}}$ ,  $\alpha_{\text{FOV}}$  and  $\alpha_{\text{GAN}}$  are loss weights for each of their respective loss components, and  $\lambda$  is the same WGAN-GP constant of Eq. (4). We set these hyperparameters to the values recommended by Yu et al.<sup>9</sup> of 1.2, 1.2, 0.001, and 10, respectively. Note that these  $L^1$  losses also utilize the *spatial-discounting* weighting for pixels within the masked region of Yu et al.<sup>9</sup>, where the weight multiplying a given pixel value within the loss is  $0.99^l$ , with  $l$  being the distance of the given pixel to the nearest *known* pixel outside of the mask. In effect, this is meant to account for the intuitive lesser ambiguity of pixels near the mask boundary than that of pixels near the mask center.





**Figure 4.** Heatmap generation with sliding window for DMSE metric.

In the training phase, patches of size  $256 \times 256$  or  $512 \times 512$  pixels were sampled randomly from the original images. Then, each patch was covered with a square-shaped mask of pre-determined side length ranging from 16 to 128 pixels. The mask was applied to a random position within the patch field of view. The patch cropping process was conducted in a way that guaranteed overlap of the patches with breast tissue, which was achieved by thresholding non-zero pixels within the random patch choice.

The model was trained with the Adam optimizer<sup>18</sup> for 2,000,000 iterations and learning rate of 0.0001 with a batch size of 9. The parameters were chosen empirically based on results on the validation set.

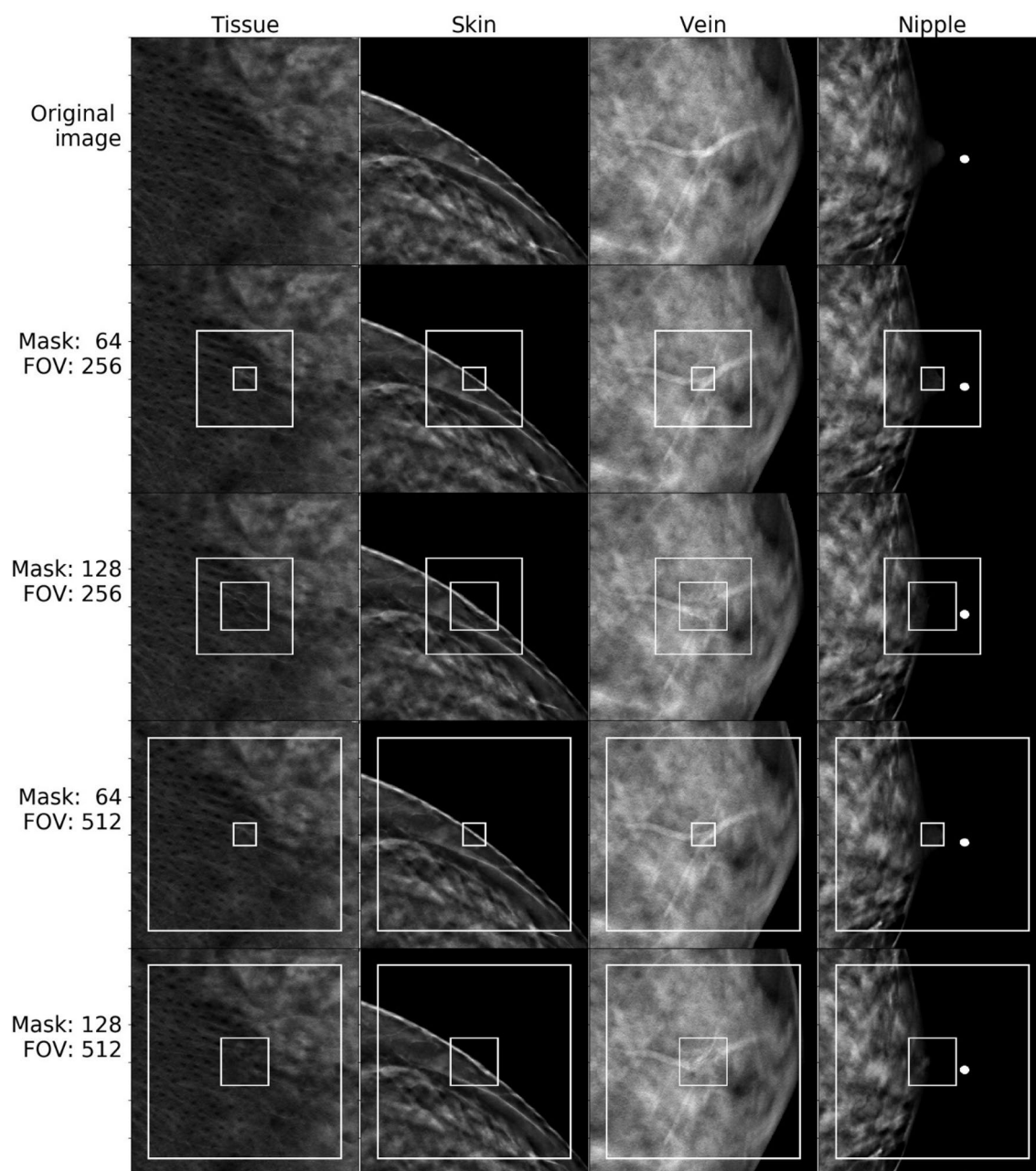
**Measuring quality of image completion for a single patch.** Once a patch is removed and inpainted by our network, one needs to assess the quality of the replacement. While different metrics could be constructed for this, we relied on two. The first was mean squared error (MSE) and the other was the discriminator loss. The discriminator loss describes the consistency of a filled region within the context of the entire input image according to the discriminator network in the GAN (Fig. 1), but its absolute value cannot be compared between different models, datasets, and stages of the model training process. However, a given model can be applied to different images to assess and compare how much a completed image resembles the data that was used to train it (a normal DBT image in this case). Finally, we also measure the product of the MSE and discriminator loss. We will refer to this metric in the further part of this paper as *DMSE*.

**Identifying abnormalities by measuring image completion quality across entire images.** Our hypothesis was that, given some test image, the locations/regions for which our algorithm have more difficulty with correctly completing are more likely to contain an abnormality. As such, to attempt to discover abnormalities within some image, we repeatedly remove parts of the image, inpaint them using our network, and measure the error across the entire image. Specifically, to measure the quality of image completion we used a sliding window approach with a shift value equal to 8 pixels. With every shift we (1) extracted a patch from the original image based on the current position of the window, (2) masked the center part of the extracted patch, (3) generated the missing part of the patch, and (4) measured and saved the computed error metric (MSE, discriminator loss, or DMSE) in a corresponding place on the abnormality heatmap. Figure 4 demonstrates the process of heatmap generation for the DMSE metric.

In addition, we also computed *averaged* heatmaps, described as follows. The process starts with creating a heatmap of the size of the original image, filled with zeros. After computing the loss for some patch, instead of saving it as a single value, we add the loss value to each pixel included within the patch to the corresponding location in the output heatmap. Because the slicing window can cover the same pixels multiple times, the pixels in the final output are divided by the number of times that were included, hence our referring to the heatmaps as “averaged”.

**Evaluation of image completion in the context of detecting abnormalities.** After generating heatmaps for every example in the test set, we measured averaged errors for locations inside and outside of radiologist-provided bounding boxes that indicate abnormalities within the set. If our approach is able to discriminate abnormal locations from normal ones, the image completion error will be notably higher for locations which include abnormalities than for normal locations. We consider only pixels inside of breast tissue (excluding background pixels with intensity value equal to zero). Also, since only the middle part of a patch is masked for completion (to ensure sufficient context for the model), areas of the input images close to the edges are not represented in the generated heatmap. The extent of this padding area depends on the field of view and mask size.

We note that although our method was only evaluated on test set images with known lesions, the lesions usually only comprised a small section of each test image. As such, the non-lesion surrounding regions of the



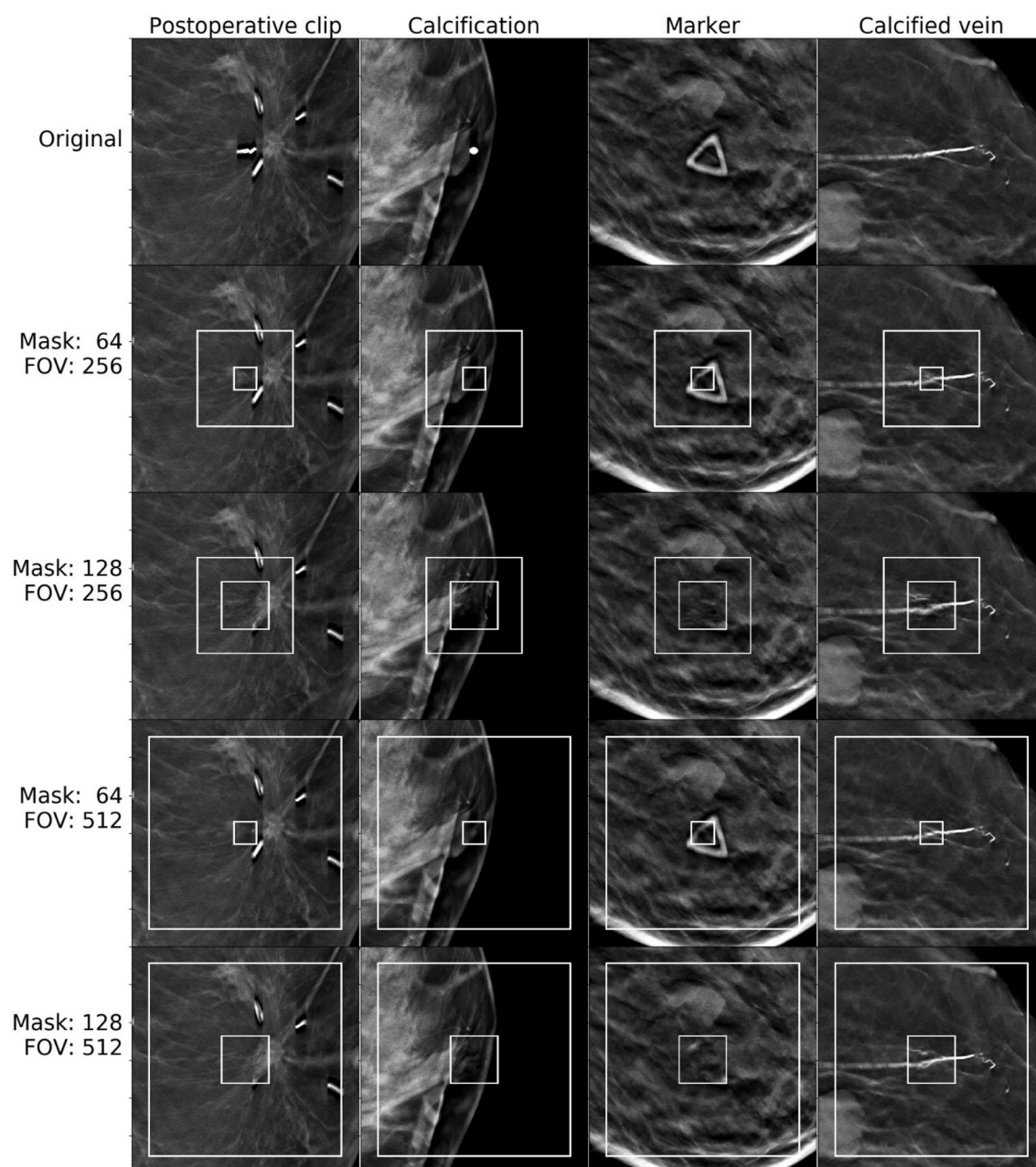
**Figure 5.** Image completion results for patches containing tissue, breast skin, veins, and nipples.

test images are just normal breast tissue, and in this way, our algorithm was tested on both normal and cancerous tissue.

## Results

**Image completion.** We provide visual references to compare image completion quality between different masks and model input sizes (fields of view, or FOV) for images from the normal set (Fig. 5) and cancer set (Fig. 6 and 7). Masks and fields of view are marked on the images as smaller and larger rectangles respectively. The part of the image covered by the mask was completed based on the remaining part of the FOV patch.

One can see that our approach is capable of generating realistic completions of breast tomosynthesis images including objects such as veins. However, once the removed patch becomes larger, the fidelity of the reconstructed object decreases. In images with unusual objects that are not lesions, when the removed patch fully covers the unusual object, the network did not accurately reconstruct the removed part. As expected, it replaced them with normal-looking tissue (Fig. 6). However, if part of the unusual object (such as a skin marker) was included in the field of view and outside of the removed patch, the network reconstructed the unusual object fairly accurately. This phenomenon was observed for normal and cancer images.



**Figure 6.** Image completion for patches containing clips, calcification, markers, and calcified veins.

**Abnormality detection.** Results for abnormality detection in terms of the mean ratio between heatmap values inside and outside of the ground truth bounding boxes and its standard deviation are given in Table 2. The table shows that the combination of MSE and discriminator loss (DMSE) outperforms the individual metrics, whereas MSE performed better than the discriminator loss. The highest value was obtained for the field of view of  $256 \times 256$  pixels with the mask size of  $128 \times 128$  pixels.

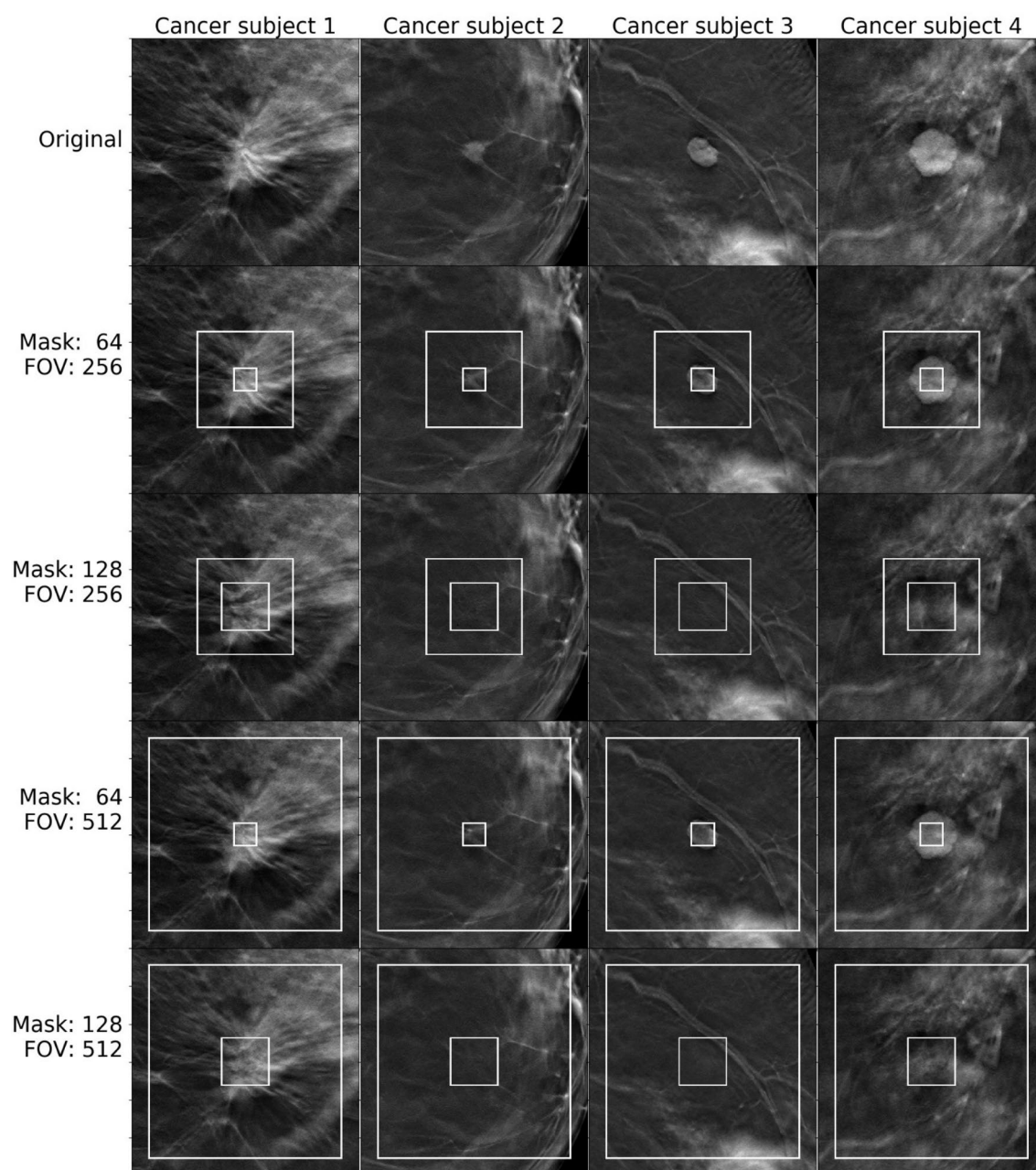
Figures 8 and 9 contain visualizations of non-averaged and averaged heatmaps, respectively, for the same subject with three separate cancer masses marked by bounding boxes. For the presented examples, the combined error metric diminishes error in areas outside of the bounding boxes as compared to error measures based solely on either MSE or discriminator loss.

## Conclusions and discussion

In this study, we used deep learning-based image completion to identify abnormal locations in digital breast tomosynthesis images. The topic is of high importance because for mammographic cancer detection—as well as many other medical imaging tasks—the availability of abnormal images is very limited and as such, an efficient use of abundantly available normal cases is crucial.

We obtained very realistic results in terms of image completion in DBT images. We showed that the trained model is able to reproduce structures like fibroglandular tissue, skin, and vessels. The covered part was completed





**Figure 7.** Image completion results for patches containing cancerous masses.

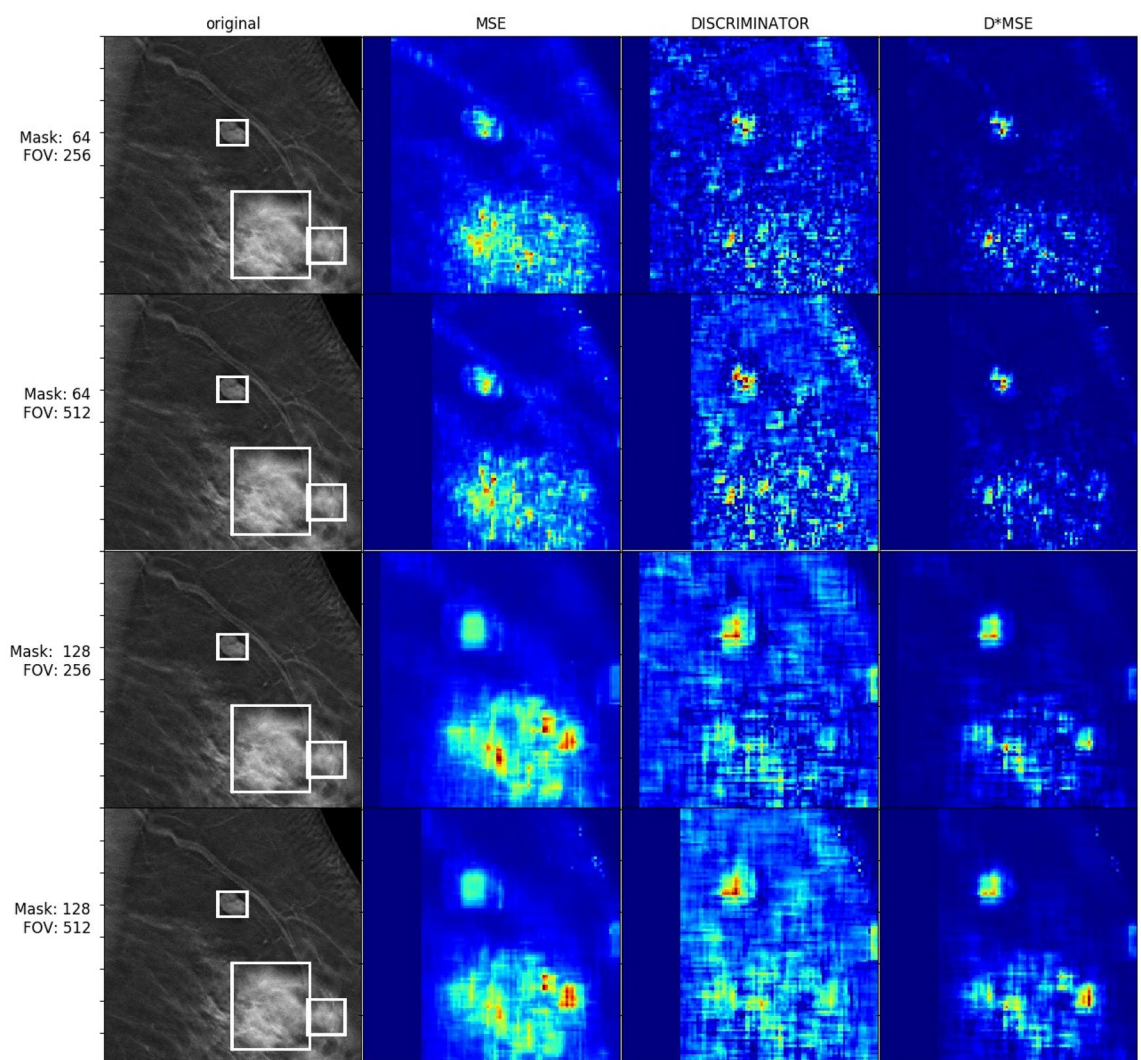
with a likely patch from a model based on its surroundings in normal locations and it does not generate unusual objects like postoperative clips or calcifications. When it comes to completing cancerous regions, the results depend on which part of the image in question is masked and completed. If a major part of a cancerous mass or architectural distortion is not covered, the model may reconstruct an abnormality but still produce loss higher than average. We have shown that there is a possibility that this approach can be used in abnormality detection.

In our experiments, we used MSE and discriminator losses in order to describe how well the image was completed. Based on our observations, MSE gives high error values for abnormal objects, e.g. post-operation clips, but also for normal tissue with complex structures such as nipples. On the other hand, we observed that discriminator loss is small while completing all kinds of shapes which were present in the training set, including nipples. Unfortunately, modest values from discriminator loss for completing parts of abnormal images make it difficult to clearly classify tissue as normal or abnormal based on that metric. The results from our study have shown that the combined loss of MSE and discriminator loss worked best. This metric gave high loss value to abnormal patches without being sensitive to sophisticated shapes present in the training set of normal cases.

From the mean ratio values of Table 2 (as well as Figs. 8 and 9), it is clear that the generator cannot inpaint/predict masks over cancerous regions nearly as accurately as that over normal breast tissue. This is to be expected, because the GAN was trained on thousands of scans of normal breast tissue, yet never saw any abnormalities

Loss type	Field of view size [pixels]	Mask size [pixels]	Mean ratio [Std]
MSE	256 × 256	64 × 64	1.93 (0.87)
MSE	256 × 256	128 × 128	2.11 (1.01)
MSE	512 × 512	64 × 64	1.83 (1.19)
MSE	512 × 512	128 × 128	1.86 (1.12)
DISCR	256 × 256	64 × 64	1.47 (0.38)
DISCR	256 × 256	128 × 128	1.46 (0.34)
DISCR	512 × 512	64 × 64	1.48 (0.63)
DISCR	512 × 512	128 × 128	1.47 (0.65)
DMSE	256 × 256	64 × 64	2.54 (2.92)
<b>DMSE</b>	<b>256 × 256</b>	<b>128 × 128</b>	<b>2.77 (1.79)</b>
DMSE	512 × 512	64 × 64	2.23 (4.33)
DMSE	512 × 512	128 × 128	2.11 (2.68)

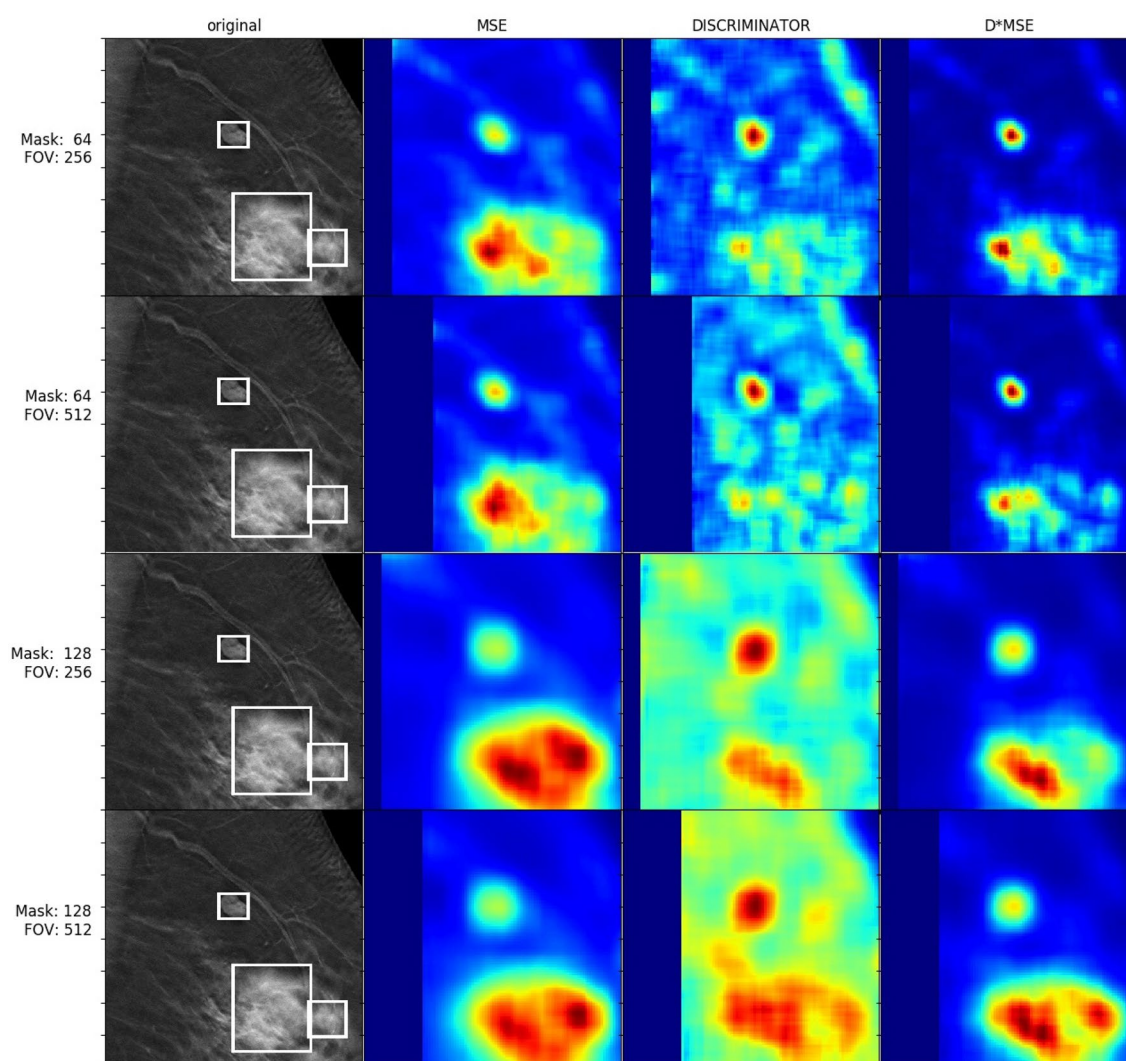
**Table 2.** Ratio of measured losses inside and outside bounding boxes for non-averaged heatmaps; DISCR = discriminator loss, std = standard deviation.



**Figure 8.** Non-averaged heatmaps for a patient with cancerous masses.

(besides the aforementioned unusual benign objects such as post-operative clips) to learn from and generalize to. Sampling from the tissue image distribution that the generator is attempting to approximate should only result in the generation of normal tissue, so it is unsurprising that the generator has difficulty with synthesizing





**Figure 9.** Averaged heatmaps for a patient with cancerous masses.

cancerous tissue, even if the surrounding context of the scan is that of a cancerous breast. Conversely, we see that when our model is used to reconstruct normal regions of breast scans that *do not* include any cancerous tissue/bounding boxes (the majority of the shown test images, as the lesions are small compared to the scale of the entire scans), the DMSE loss (our best-performing metric) is noticeably smaller than in the cancerous regions, on average, showing that normal regions can still be differentiated from cancerous within the output heatmaps. This is because the DMSE loss is proportional to the discriminator loss, which describes how real the discriminator judges the generated tissue to be, compared to the ground truth tissue. Because the discriminator is only trained to discriminate between realistic and non-realistic *normal* tissue, and the generator can only perform poorly and unrealistically when reconstructing regions with abnormalities, this loss is indeed greater for such cancerous regions. As such, the usage of the discriminator provides a further refinement to the loss metrics that are used to indicate abnormalities, explaining why the DMSE metric performed best.

We note that in order to provide classification metrics for our method such as accuracy or ROC (receiver operating characteristic) curves, we would first need to define exactly what a “detection” of cancerous/abnormal tissue is in the context of the generated heatmaps (Figs. 8 and 9). This would require choosing some numerical threshold for the pixel values of the inpainting error/DMSE (Table 2), such that if the error for some pixels/region was greater than this threshold, this region could be described as being detected to be abnormal. In turn, we could compare these error values to known lesion bounding boxes (or use the ratio of error between inside and outside the boxes, as in Table 2) to obtain metrics such as the true positive rate (TPR). However, doing this properly would require further research and experimentation, including questions of the definition of detected regions, overlap criteria, postprocessing and false positive reduction and other questions that we believe are beyond the scope of this work which focuses on the concept of image inpainting.

GANs have proven to be useful in a range of applications, including realistic facial image creation and customization (e.g.<sup>19</sup>, image-to-image translation (e.g.<sup>20</sup>, and even lesser-known applications such as excising rain from images<sup>21</sup> among others. More particularly related to our method, GAN-based inpainting itself has also seen wide use for a variety of applications, from the conversion of 2D images to 3D representations<sup>22</sup>, to temporally

consistent video completion/inpainting<sup>23</sup>, to automatic face-anonymizing for privacy<sup>24</sup>. The use of GANs for abnormality detection is not nearly as common as the aforementioned trend of using GANs for other purposes. However, works such as Herent et al.<sup>25</sup>, Cao et al.<sup>26</sup>, Kooi et al.<sup>27</sup>, Yap et al.<sup>28</sup> and Yap et al.<sup>29</sup> also use deep learning for breast lesion detection (e.g., lesion type classification, object recognition and/or segmentation), but they rely on the direct, supervised learning of the appearance of real breast lesions, and as such are distinctly different from our semi-supervised, normal data-based GAN method. Despite this, there is still a group of other generative modeling-based lesion/abnormality detection methods that can be compared to ours.

Benson & Beets-Tan<sup>30</sup> introduced a method that uses GANs (but with a different inpainting algorithm) to learn the data distribution of normal brain scans and perform inpainting on a grid of masks over the input image, like our method. In this experiment, the sum of the pixel-wise inpainting residuals within each mask are used to indicate abnormalities *mask-by-mask* (if this sum is above a certain numerical threshold), this is essentially the same as our method, just with slightly different masking techniques that create the final outputted abnormality heatmap. Li et al.<sup>31</sup> proposed a method that is also similar in practice to ours and that of Benson & Beets-Tan, insofar that input test images are divided into mask regions, which are then each separately inpainted one-by-one, after which an anomaly heatmap is generated according to the discrepancy between the original image and the reconstructed image. This model, although originally designed for visual anomaly detection in the context of industrial inspection, is essentially the same idea as our method, with the one difference being that it utilizes encoder-based, rather than the more advanced GAN-based inpainting used in our study.

The advantage of our model is that we do not have to rely on limited cancerous image data. Instead, we train on the abundance of normal scans, a philosophy that certain similar studies share. Chen et al.<sup>13</sup> also uses an adversarial (but auto-encoder) approach to learning the distribution of healthy *brain* tissue, by learning the mapping of the input scan to some latent space and detecting anomalous scans within this space itself. Schlegl et al.<sup>10</sup> similarly uses a DCGAN-based architecture (as well as an encoder) to learn the latent space representation and generative process for normal anatomical image data such that at test time, unseen images are mapped to this latent space, and if anomalous, will be noticeably different from their reconstruction, which is found by mapping back from the latent space representation. From here, anomalous regions within the input are detected based on this discrepancy, including both reconstruction and discriminative losses. These methods are similar to ours because of the training on non-anomalous data to learn the distribution for such data. However, there are two main differences when compared to our work. The first is that these works perform image reconstruction using latent representations of data, not with inpainting/direct image completion. The second difference is that at test time, our method performs reconstruction of some masked region using the surrounding non-masked region as input to the network (not viewing the covered region to be reconstructed), while these methods are applied to reconstruct entire images, not patches, of which the networks use the *entire image* as input, not excluding anything to be used in inference, which is distinctly different than our method.

Just as our model does, the two methods of the last paragraph can be used to produce abnormality heatmaps similar to Figs. 8 and 9. However, it is important to note that in the case of the first, auto-encoder-based model, the authors state that the reconstruction quality is predicated on the input image being downsampled to a  $32 \times 32$  resolution. Doing such for our data would drastically reduce the quality of our very high-resolution DBT scan images (even in the training phase, as this uses  $256 \times 256$  inputs), which could produce unforeseen consequences within the training procedure and testing inference. Therefore, in order to compare this method to ours, we would have to use it in a way that it was not intended or downsample our data by a factor of 64 which would dramatically degrade its quality. Similarly, the second method (f-AnoGAN) is built with a DCGAN/WGAN architecture that is designed to have stable training specifically for  $64 \times 64$  images. While this resolution is greater than  $32 \times 32$ , either drastically downsampling our input to this resolution, or augmenting the network to accept a larger input, could produce unwanted training issues, or test inference/heatmaps that are not necessarily valid to compare with ours. Alternatively, one could imagine using these methods along a “grid” of disjoint partitions of the input test image, to preserve global test image resolution, but this could potentially result in issues with global cross-partition coherence and consistency. In summary, directly comparing our model to these two models, which are built for lower-resolution images, would likely require considerable further research and development before we obtained results that we are confident in, and as such, this is also beyond the scope of this proof-of-concept work. This reason and the argument outlined in the previous paragraph are the main points for why our method is not immediately reasonable to quantitatively compare to other techniques; in other words, methods with which we can reasonably compare to do not exist.

Our study has certain limitations. First, the number of positive cases in the test set was small. However, it was sufficient to provide a good overview of the algorithm’s performance, on both normal tissue (image regions without lesions) and cancerous tissue. Second, the dataset used for evaluation did not contain annotations for all kinds of abnormal objects, e.g. post-operation clips, which did not allow us to provide detailed performance estimation of detection quality. Moreover, the range of tested sizes for the field of view parameter was limited to what we considered reasonable and computationally feasible but larger range of values could be considered in future studies. Another limitation of our approach is that it identified unusual locations in the images that were *not* cancerous. This could be potentially addressed by oversampling such structures during the training. Our approach was also limited by computation time to generate heatmaps since those required thousands of model runs per image. Finally, our no-padding approach leads to omitting boundary parts of the image during detection.

In summary, we showed promising results on how to effectively use data without objects of interest for detection of abnormalities in medical images. Our approach could be further refined via a number of approaches, such as by combining it with fully supervised methods in order to improve performance of object detection with a scarce training signal.



Received: 2 July 2020; Accepted: 20 April 2021  
Published online: 13 May 2021

## References

1. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* <https://doi.org/10.1016/j.media.2017.07.005> (2017).
2. Lehman, C. D. *et al.* National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology* **283**(1), 49–58 (2017).
3. Gilbert, F. J., Tucker, L. & Young, K. C. Digital breast tomosynthesis (DBT): a review of the evidence for use as a screening tool. *Clin. Radiol.* **71**(2), 141–150 (2016).
4. Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **106**, 249–259. <https://doi.org/10.1016/j.neunet.2018.07.011> (2018).
5. Masci, J., Meier, U., Cireşan, D. & Schmidhuber, J. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6791 LNCS, pp. 52–59). (Springer, 2011). [https://doi.org/10.1007/978-3-642-21735-7\\_7](https://doi.org/10.1007/978-3-642-21735-7_7)
6. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P. A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning* 1096–1103. (ACM Press, New York, 2008). <https://doi.org/10.1145/1390156.1390294>
7. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. & Bengio, Y. Generative Adversarial Nets. In *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2672–2680 (2014).
8. Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016—Conference Track Proceedings*. International Conference on Learning Representations, ICLR (2016).
9. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X. & Huang, T. S. Generative Image Inpainting with Contextual Attention. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 5505–5514 (2018).
10. Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G. & Schmidt-Erfurth, U. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Med. Image Anal.* **54**, 30–44. <https://doi.org/10.1016/j.media.2019.01.010> (2019).
11. Haselmann, M., Gruber, D. P. & Tabatabai, P. Anomaly detection using deep learning based image completion. In *Proceedings—17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018* 1237–1242. Institute of Electrical and Electronics Engineers Inc. (2019). <https://doi.org/10.1109/ICMLA.2018.00201>
12. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation (2017). [arXiv:1710.10196](https://arxiv.org/abs/1710.10196).
13. Chen, J., Chen, J., Chao, H. & Yang, M. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3155–3164 (2018).
14. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* 2223–2232 (2017).
15. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. & Frey, B. Adversarial autoencoders (2015). [arXiv:1511.05644](https://arxiv.org/abs/1511.05644).
16. Arjovsky, M., Chintala, S., & Bottou, L. Wasserstein GAN (2017).
17. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. C. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems* 5767–5777 (2017).
18. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization (2014). [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
19. Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I. & Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets (2016). [arXiv:1606.03657](https://arxiv.org/abs/1606.03657).
20. Yi, Z., Zhang, H., Tan, P. & Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision* 2849–2857 (2017).
21. Zhang, H., Sindagi, V. & Patel, V. M. (2017). Image de-raining using a conditional generative adversarial network. [arXiv:1701.05957](https://arxiv.org/abs/1701.05957).
22. Shih, M. L., Su, S. Y., Kopf, J. & Huang, J. B. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8028–8038 (2016).
23. Xu, R., Li, X., Zhou, B. & Loy, C. C. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3723–3732 (2019).
24. Hukkelås, H., Mester, R., Lindseth, F. Deepprivacy: a generative adversarial network for face anonymization. In *International Symposium on Visual Computing* 565–578. (Springer, 2019)
25. Herent, P. *et al.* Detection and characterization of MRI breast lesions using deep learning. *Diagn. Interv. Imaging* **100**(4), 219–225 (2019).
26. Cao, Z., Duan, L., Yang, G., Yue, T. & Chen, Q. An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. *BMC Med. Imaging* **19**(1), 1–9 (2019).
27. Kooi, T. *et al.* Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017).
28. Yap, M. H. *et al.* Breast ultrasound region of interest detection and lesion localisation. *Artif. Intell. Med.* **107**, 101880 (2020).
29. Yap, M. H. *et al.* Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE j. Biomed. Health Inform.* **22**(4), 1218–1226 (2017).
30. Benson, S. & Beets-Tan, R. GAN-based anomaly detection in multi-modal MRI images. *bioRxiv* (2020).
31. Li, Z., Li, N., Jiang, K., Ma, Z., Wei, X., Hong, X. & Gong, Y. Superpixel Masking and Inpainting for Self-Supervised Anomaly Detection (2020).

## Author contributions

A.S prepared software used in the publication, generated results and figures. M.B prepared data used in the publication, revised the study and results. M.A.M. supervised the study. All authors wrote and reviewed the manuscript.

## Funding

National Institute of Biomedical Imaging and Bioengineering, Grand No. 1R01EB021360.

## Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to A.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

## **Appendix B**

### **Authorship statements**

Maciej A Mazurowski, PhD  
Department of Radiology  
Duke University

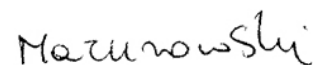
### **Authorship Statement**

I hereby declare that my contributions to the paper

Maciej A Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R Bashir. “Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI.” *Journal of Magnetic Resonance Imaging*, 49(4):939–954, 2019.

included:

- conceptualization,
- methodology,
- investigation,
- resources,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization,
- supervision,
- project administration.



Maciej A Mazurowski

Warsaw, Poland, 22 December 2022

Mateusz Buda, MSc  
Department of Radiology  
Duke University

### **Authorship Statement**

I herby declare that my contributions to the paper

Maciej A Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R Bashir. "Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI." *Journal of Magnetic Resonance Imaging*, 49(4):939–954, 2019.

included:

- investigation,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization.

Mateusz Buda

Mateusz Buda

Hamilton, Ontario, Canada, 26 December 2022

Ashirbani Saha, PhD  
Department of Oncology  
McMaster University  
Hamilton, ON, Canada  
(Formerly, Department of Radiology  
Duke University, USA)

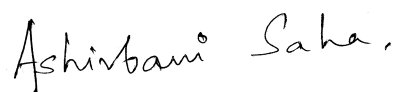
### Authorship Statement

I hereby declare that my contributions to the paper

Maciej A. Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R. Bashir.  
“Deep learning in radiology: an overview of the concepts and a survey of the  
state of the art with focus on MRI.” Journal of Magnetic Resonance Imaging,  
49(4):939–954, 2019.

included:

- conceptualization,
- methodology,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization.



---

Ashirbani Saha



Mustafa R Bashir, MD  
Department of Radiology  
Duke University

#### Authorship Statement

I hereby declare that my contributions to the paper

Maciej A Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R Bashir.  
“Deep learning in radiology: an overview of the concepts and a survey of the  
state of the art with focus on MRI.” *Journal of Magnetic Resonance Imaging*,  
49(4):939–954, 2019.

included:

- conceptualization,
- methodology,
- validation,
- formal analysis,
- investigation,
- resources,
- data curation,
- writing - original draft,
- writing - review & editing,
- supervision,
- project administration.

---

Mustafa R Bashir

Warsaw, Poland, 22 December 2022

Mateusz Buda, MSc  
Department of Radiology  
Duke University

### **Authorship Statement**

I hereby declare that my contributions to the paper

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks." Neural Networks, 106:249–259, 2018.

included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization,
- project administration.

Mateusz Buda

Mateusz Buda

Stockholm, Sweden, 06 February 2023

Atsuto Maki, Professor  
School of Electrical Engineering and Computer Science (EECS)  
KTH Royal Institute of Technology

### Authorship Statement

I herby declare that my contributions to the paper

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks." *Neural Networks*, 106:249–259, 2018.

included:

- conceptualization,
- methodology,
- validation,
- investigation,
- writing - review & editing,
- supervision,
- project administration.



---

Atsuto Maki

Maciej A Mazurowski, PhD  
Department of Radiology  
Duke University

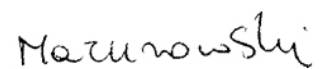
### **Authorship Statement**

I hereby declare that my contributions to the paper

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks." *Neural Networks*, 106:249–259, 2018.

included:

- conceptualization,
- methodology,
- validation,
- investigation,
- resources,
- writing - review & editing,
- visualization,
- supervision,
- project administration,
- funding acquisition.



Maciej A Mazurowski

Warsaw, Poland, 22 December 2022

Mateusz Buda, MSc  
Department of Radiology  
Duke University

### **Authorship Statement**

I herby declare that my contributions to the paper

Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm." *Computers in Biology & Medicine*, 109:218–225, 2019.

included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization.

Mateusz Buda

Mateusz Buda

Hamilton, Ontario, Canada, 26 December 2022

Ashirbani Saha, PhD  
Department of Oncology  
McMaster University  
Hamilton, ON, Canada  
(Formerly, Department of Radiology  
Duke University, USA)

#### Authorship Statement

I hereby declare that my contributions to the paper

Mateusz Buda, Ashirbani Saha, and Maciej A. Mazurowski. "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm." *Computers in Biology & Medicine*, 109:218–225, 2019.

included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - review & editing.



---

Ashirbani Saha



Maciej A Mazurowski, PhD  
Department of Radiology  
Duke University

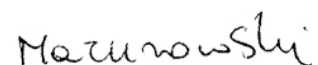
### **Authorship Statement**

I hereby declare that my contributions to the paper

Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. “Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm.” *Computers in Biology & Medicine*, 109:218–225, 2019.

included:

- conceptualization,
- methodology,
- validation,
- formal analysis,
- investigation,
- resources,
- writing - review & editing,
- visualization,
- supervision,
- project administration,
- funding acquisition.



Maciej A Mazurowski

Warsaw, Poland, 22 December 2022

Mateusz Buda, MSc  
Department of Radiology  
Duke University

### **Authorship Statement**

I herby declare that my contributions to the paper

Mateusz Buda, Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. "Deep radiogenomics of lower-grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images." *Radiology: Artificial Intelligence*, 2(1), 2020.

included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization.

Mateusz Buda

Mateusz Buda

Albany, NY, USA, 22 December 2022

Ehab A AlBadawy, BSc  
Department of Radiology  
Duke University

### Authorship Statement

I hereby declare that my contributions to the paper

Mateusz Buda, Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski.  
“Deep radiogenomics of lower-grade gliomas: convolutional neural networks  
predict tumor genomic subtypes using MR images.” Radiology: Artificial  
Intelligence, 2(1), 2020.

included:

- conceptualization,
- methodology,
- software,
- data curation,
- writing - review & editing.

DocuSigned by:  
*Ehab A AlBadawy*  
F5978626103F421...

Ehab A AlBadawy

Hamilton, Ontario, Canada, 26 December 2022

Ashirbani Saha, PhD  
Department of Oncology  
McMaster University  
Hamilton, ON, Canada  
(Formerly, Department of Radiology  
Duke University, USA)

### Authorship Statement

I hereby declare that my contributions to the paper

Mateusz Buda, Ehab A. AlBadawy, Ashirbani Saha, and Maciej A. Mazurowski.  
“Deep radiogenomics of lower-grade gliomas: convolutional neural networks  
predict tumor genomic subtypes using MR images.” Radiology: Artificial  
Intelligence, 2(1), 2020.

included:

- methodology,
- validation,
- investigation,
- data curation,
- writing - review & editing.



---

Ashirbani Saha

Maciej A Mazurowski, PhD  
Department of Radiology  
Duke University

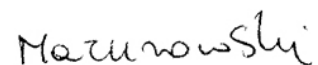
### **Authorship Statement**

I herby declare that my contributions to the paper

Mateusz Buda, Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. "Deep radiogenomics of lower-grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images." Radiology: Artificial Intelligence, 2(1), 2020.

included:

- conceptualization,
- methodology,
- validation,
- investigation,
- resources,
- writing - review & editing,
- supervision,
- project administration,
- funding acquisition.



Maciej A Mazurowski

Warsaw, Poland, 22 December 2022

Mateusz Buda, MSc  
Department of Radiology  
Duke University

### **Authorship Statement**

I hereby declare that my contributions to the paper

Mateusz Buda, Benjamin Wildman-Tobriner, Kerry Castor, Jenny K Hoang, and Maciej A Mazurowski. "Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images." *Ultrasound in Medicine & Biology*, 46(2):415–421, 2020.

included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization.

Mateusz Buda

Mateusz Buda

Benjamin Marshall Wildman-Tobriner, MD  
Department of Radiology  
Duke University

### **Authorship Statement**

I herby declare that my contributions to the paper

Mateusz Buda, Benjamin Wildman-Tobriner, Kerry Castor, Jenny K Hoang, and Maciej A Mazurowski. “Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images.” *Ultrasound in Medicine & Biology*, 46(2):415–421, 2020.

included:

- conceptualization,
- methodology,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing.



---

Benjamin Marshall Wildman-Tobriner



Kerry Castor, BSc  
Department of Radiology  
Duke University

### Authorship Statement

I hereby declare that my contributions to the paper

Mateusz Buda, Benjamin Wildman-Tobriner, Kerry Castor, Jenny K Hoang, and Maciej A Mazurowski. "Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images." *Ultrasound in Medicine & Biology*, 46(2):415–421, 2020.

included:

- methodology,
- software,
- formal analysis,
- investigation,
- data curation,
- writing - review & editing,
- visualization.

A handwritten signature in black ink that reads "Kerry Castor". The signature is written in a cursive style and is positioned above a horizontal line.

Kerry Castor

Baltimore, MD, USA, 22 December 2022

Jenny K Hoang, MBBS, MBS, MHS  
Department of Radiology  
Duke University

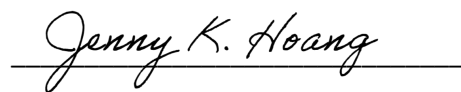
#### Authorship Statement

I hereby declare that my contributions to the paper

Mateusz Buda, Benjamin Wildman-Tobriner, Kerry Castor, Jenny K Hoang, and Maciej A Mazurowski. "Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images." *Ultrasound in Medicine & Biology*, 46(2):415–421, 2020.

included:

- conceptualization,
- methodology,
- validation,
- investigation,
- writing - review & editing,
- supervision,
- project administration.

A handwritten signature in cursive script that reads "Jenny K. Hoang". The signature is written in black ink and is positioned above a horizontal line.

Jenny K Hoang

Maciej A Mazurowski, PhD  
Department of Radiology  
Duke University

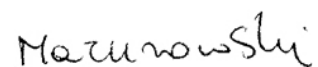
### **Authorship Statement**

I hereby declare that my contributions to the paper

Mateusz Buda, Benjamin Wildman-Tobriner, Kerry Castor, Jenny K Hoang, and Maciej A Mazurowski. "Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images." *Ultrasound in Medicine & Biology*, 46(2):415–421, 2020.

included:

- conceptualization,
- methodology,
- validation,
- investigation,
- resources,
- data curation,
- writing - review & editing,
- supervision,
- project administration,
- funding acquisition.



Maciej A Mazurowski

Warsaw, Poland, 22 December 2022

Mateusz Buda, MSc  
Department of Radiology  
Duke University

### Authorship Statement

I hereby declare that my contributions to the paper

Mateusz Buda, Benjamin Wildman-Tobriner, Jenny K Hoang, David Thayer, Franklin N Tessler, William D Middleton, and Maciej A Mazurowski. "Management of thyroid nodules seen on US images: deep learning may match performance of radiologists." Radiology, 292(3):695–701, 2019.

included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization.

Mateusz Buda

Mateusz Buda

Benjamin Marshall Wildman-Tobriner, MD  
Department of Radiology  
Duke University


### **Authorship Statement**

I herby declare that my contributions to the paper

Mateusz Buda, Benjamin Wildman-Tobriner, Jenny K Hoang, David Thayer, Franklin N Tessler, William D Middleton, and Maciej A Mazurowski. "Management of thyroid nodules seen on US images: deep learning may match performance of radiologists." Radiology, 292(3):695–701, 2019.

included:

- conceptualization,
- methodology,
- validation,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing.



---

Benjamin Marshall Wildman-Tobriner

Baltimore, MD, USA, 22 December 2022

Jenny K Hoang, MBBS, MBS, MHS  
Department of Radiology  
Duke University

#### Authorship Statement

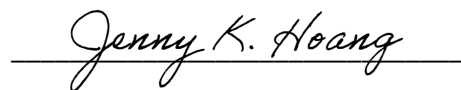
I hereby declare that my contributions to the paper

Mateusz Buda, Benjamin Wildman-Tobriner, Jenny K Hoang, David Thayer,  
Franklin N Tessler, William D Middleton, and Maciej A Mazurowski.

“Management of thyroid nodules seen on US images: deep learning may match  
performance of radiologists.” *Radiology*, 292(3):695–701, 2019.

included:

- conceptualization,
- validation,
- investigation,
- writing - review & editing,
- visualization,
- supervision,
- project administration.

  
Jenny K Hoang

St Louis, MO, USA, 22 December 2022

William D Middleton, MD  
Mallinckrodt Institute of Radiology  
Washington University in St. Louis

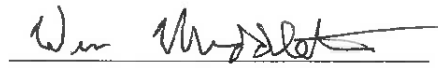
#### Authorship Statement

I hereby declare that my contributions to the paper

Mateusz Buda, Benjamin Wildman-Tobriner, Jenny K Hoang, David Thayer,  
Franklin N Tessler, William D Middleton, and Maciej A Mazurowski.  
“Management of thyroid nodules seen on US images: deep learning may match  
performance of radiologists.” Radiology, 292(3):695–701, 2019.

included:

- conceptualization,
- validation,
- investigation,
- resources,
- data curation,
- writing - review & editing.



William D Middleton



Maciej A Mazurowski, PhD  
Department of Radiology  
Duke University

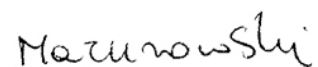
### **Authorship Statement**

I herby declare that my contributions to the paper

Mateusz Buda, Benjamin Wildman-Tobriner, Jenny K Hoang, David Thayer, Franklin N Tessler, William D Middleton, and Maciej A Mazurowski. "Management of thyroid nodules seen on US images: deep learning may match performance of radiologists." Radiology, 292(3):695–701, 2019.

included:

- conceptualization,
- methodology,
- validation,
- formal analysis,
- investigation,
- resources,
- writing - review & editing,
- supervision,
- project administration,
- funding acquisition.



Maciej A Mazurowski

Benjamin Marshall Wildman-Tobriner, MD  
Department of Radiology  
Duke University

### **Authorship Statement**

I herby declare that my contributions to the paper

Benjamin Wildman-Tobriner, Mateusz Buda, Jenny K Hoang, William D Middleton, David Thayer, Ryan G Short, Franklin N Tessler, and Maciej A Mazurowski. "Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility." Radiology, 292(1):112–119, 2019.

included:

- conceptualization,
- methodology,
- validation,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing.



---

Benjamin Marshall Wildman-Tobriner

Warsaw, Poland, 22 December 2022

Mateusz Buda, MSc  
Department of Radiology  
Duke University

### **Authorship Statement**

I hereby declare that my contributions to the paper

Benjamin Wildman-Tobriner, Mateusz Buda, Jenny K Hoang, William D Middleton, David Thayer, Ryan G Short, Franklin N Tessler, and Maciej A Mazurowski. "Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility." *Radiology*, 292(1):112–119, 2019.

included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization.

Mateusz Buda

Mateusz Buda

Baltimore, MD, USA, 22 December 2022

Jenny K Hoang, MBBS, MBS, MHS  
Department of Radiology  
Duke University

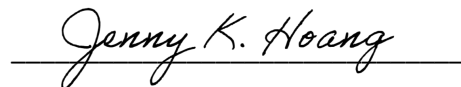
#### Authorship Statement

I hereby declare that my contributions to the paper

Benjamin Wildman-Tobriner, Mateusz Buda, Jenny K Hoang, William D Middleton, David Thayer, Ryan G Short, Franklin N Tessler, and Maciej A Mazurowski. "Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility." *Radiology*, 292(1):112–119, 2019.

included:

- conceptualization,
- validation,
- investigation,
- resources,
- writing - review & editing,
- supervision,
- project administration.

A handwritten signature in cursive script that reads "Jenny K. Hoang". The signature is written in black ink and is positioned above a horizontal line.

Jenny K Hoang

St Louis, MO, USA, 22 December 2022

William D Middleton, MD  
Mallinckrodt Institute of Radiology  
Washington University in St. Louis

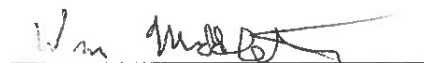
#### Authorship Statement

I hereby declare that my contributions to the paper

Benjamin Wildman-Tobriner, Mateusz Buda, Jenny K Hoang, William D Middleton, David Thayer, Ryan G Short, Franklin N Tessler, and Maciej A Mazurowski. "Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility." *Radiology*, 292(1):112–119, 2019.

included:

- conceptualization,
- validation,
- investigation,
- resources,
- data curation,
- writing - review & editing.



William D Middleton

Maciej A Mazurowski, PhD  
Department of Radiology  
Duke University

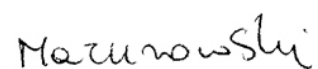
### **Authorship Statement**

I hereby declare that my contributions to the paper

Benjamin Wildman-Tobriner, Mateusz Buda, Jenny K Hoang, William D Middleton, David Thayer, Ryan G Short, Franklin N Tessler, and Maciej A Mazurowski. "Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility." Radiology, 292(1):112–119, 2019.

included:

- conceptualization,
- methodology,
- validation,
- formal analysis,
- investigation,
- resources,
- writing - review & editing,
- supervision,
- project administration,
- funding acquisition.



Maciej A Mazurowski

Warsaw, Poland, 22 December 2022

Mateusz Buda, MSc  
Department of Radiology  
Duke University

### **Authorship Statement**

I herby declare that my contributions to the paper

Mateusz Buda, Ashirbani Saha, Ruth Walsh, Sujata Ghate, Nianyi Li, Albert Świącicki, Joseph Y Lo, and Maciej A Mazurowski. "A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images." JAMA network open, 4(8), 2021.

included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization.

Mateusz Buda

Mateusz Buda



Hamilton, Ontario, Canada, 26 December 2022

Ashirbani Saha, PhD  
Department of Oncology  
McMaster University  
Hamilton, ON, Canada  
(Formerly, Department of Radiology  
Duke University, USA)

### Authorship Statement

I hereby declare that my contributions to the paper

Mateusz Buda, Ashirbani Saha, Ruth Walsh, Sujata Ghate, Nianyi Li, Albert Świącicki, Joseph Y. Lo, and Maciej A. Mazurowski. "A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images." JAMA network open, 4(8), 2021.

included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing.



---

Ashirbani Saha

Nianyi Li, PhD  
Department of Radiology  
Duke University

### Authorship Statement

I herby declare that my contributions to the paper

Mateusz Buda, Ashirbani Saha, Ruth Walsh, Sujata Ghatte, Nianyi Li, Albert Świącicki, Joseph Y Lo, and Maciej A Mazurowski. "A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images." JAMA network open, 4(8), 2021.

included:

- conceptualization,
- methodology,
- software,
- validation,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing.



Nianyi Li

Warsaw, Poland, 12/24/2022

Albert Świącicki, BSc  
Department of Radiology  
Duke University

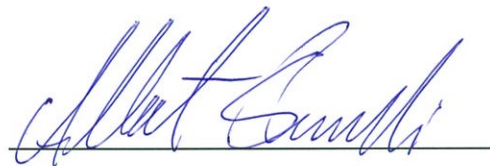
#### Authorship Statement

I hereby declare that my contributions to the paper

Mateusz Buda, Ashirbani Saha, Ruth Walsh, Sujata Ghate, Nianyi Li, Albert Świącicki, Joseph Y Lo, and Maciej A Mazurowski. "A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images." JAMA network open, 4(8), 2021.

included:

- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- writing - review & editing,
- visualization.



Albert Świącicki

Maciej A Mazurowski, PhD  
Department of Radiology  
Duke University

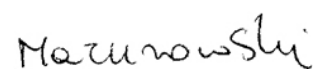
### **Authorship Statement**

I hereby declare that my contributions to the paper

Mateusz Buda, Ashirbani Saha, Ruth Walsh, Sujata Ghate, Nianyi Li, Albert Świącicki, Joseph Y Lo, and Maciej A Mazurowski. "A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images." JAMA network open, 4(8), 2021.

included:

- conceptualization,
- methodology,
- validation,
- investigation,
- resources,
- data curation,
- writing - review & editing,
- supervision,
- project administration,
- funding acquisition.



Maciej A Mazurowski

Warsaw, Poland, 12/24/2022

Albert Świącicki, BSc  
Department of Radiology  
Duke University

#### Authorship Statement

I hereby declare that my contributions to the paper

Albert Świącicki, Nicholas Konz, Mateusz Buda, and Maciej A Mazurowski. "A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis." Scientific reports, 11(1):1–13, 2021

i included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing,
- visualization.



Albert Świącicki

Durham, NC, USA, 22 December 2022

Nicholas Konz, BSc  
Department of Radiology  
Duke University

### Authorship Statement

I hereby declare that my contributions to the paper

Albert Swiecicki, Nicholas Konz, Mateusz Buda, and Maciej A Mazurowski. "A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis." Scientific reports, 11(1):1–13, 2021

included:

- methodology,
- software,
- formal analysis,
- investigation,
- writing - review & editing,
- visualization.



---

Nicholas Konz

Warsaw, Poland, 22 December 2022

Mateusz Buda, MSc  
Department of Radiology  
Duke University

### **Authorship Statement**

I herby declare that my contributions to the paper

Albert Swiecicki, Nicholas Konz, Mateusz Buda, and Maciej A  
Mazurowski. "A generative adversarial network-based abnormality  
detection using only normal images for model training with application  
to digital breast tomosynthesis." Scientific reports, 11(1):1–13, 2021

included:

- conceptualization,
- methodology,
- software,
- validation,
- formal analysis,
- investigation,
- data curation,
- writing - original draft,
- writing - review & editing.

Mateusz Buda

Mateusz Buda



Maciej A Mazurowski, PhD  
Department of Radiology  
Duke University

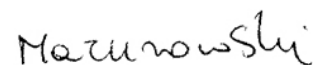
### **Authorship Statement**

I herby declare that my contributions to the paper

Albert Swiecicki, Nicholas Konz, Mateusz Buda, and Maciej A Mazurowski. "A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis." Scientific reports, 11(1):1–13, 2021

included:

- conceptualization,
- methodology,
- validation,
- investigation,
- resources,
- writing - review & editing,
- supervision,
- project administration,
- funding acquisition.



Maciej A Mazurowski