Developing Informatics Frameworks for Evaluating Deep Learning Algorithms for Mammography

Ojas Ankurbhai Ramwala

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

John H. Gennari, Chair

Christoph I. Lee

William Lotter

Sean D. Mooney

Program Authorized to Offer Degree:
Department of Biomedical Informatics and Medical Education

University of Washington

Abstract

Developing Informatics Frameworks for Evaluating Deep Learning Algorithms for Mammography

Ojas Ankurbhai Ramwala

Chair of the Supervisory Committee:
John H. Gennari
Department of Biomedical Informatics and Medical Education

Deep learning algorithms have played a major role in advancing AI for mammography-based breast cancer screening. Studies have shown that AI tools can achieve, and in some cases exceed, the performance of breast imaging radiologists. Integrating deep learning algorithms into clinical workflows has the potential to streamline mammography interpretation, aid early cancer detection, and enhance risk prediction. Nevertheless, in spite of promising initial performance across a broad range of tasks in mammography interpretation and breast cancer screening, their adoption in clinical settings remains limited. A major contributing factor is the lack of methods to thoroughly evaluate the safety, reliability, clinical utility, and trustworthiness of AI models in mammography, thereby creating a critical gap between algorithm development and real-world implementation. Therefore, we have developed informatics frameworks to support a comprehensive and systematic evaluation of AI algorithms prior to their clinical adoption, enabling stakeholders to critically assess model generalizability and interpret the underlying inference mechanism that drives model predictions.

AI models may have limited generalizability in new clinical settings or within specific demographic subpopulations. We developed an open-source framework, ClinValAI (Clinical Validation of AI), to support health systems in implementing a cloud-based infrastructure for rigorous external validation of AI algorithms before clinical adoption. ClinValAI enables secure, privacy-preserving external validation by protecting patient imaging data and

developers' intellectual property while offering scalable, customizable workflows that accommodate the diverse computational demands of multiple AI algorithms. We demonstrate ClinValAI's utility by performing a large-scale external validation of multiple FDA-cleared commercial AI algorithms for breast cancer detection using mammography exams from seven U.S. regional breast cancer registries. By comparing those algorithms and evaluating their performance against radiologists' assessments, our study highlights the benefits and risks of adopting AI tools in clinical workflows. ClinValAI provides a holistic framework for validating medical imaging models and has the potential to advance the adoption of accurate, generalizable AI models in mammography-based breast cancer screening.

Even when AI algorithms demonstrate strong generalizability, their 'black box' nature obscures meaningful insights into their inference mechanism, undermining trust and transparency. We address this challenge specifically for the Mirai model for mammography-based breast cancer risk prediction. We developed a method, FOCUS (Feature-space OCclusion for Understanding Saliency), to assess the contribution of different mammogram imaging regions to Mirai's prediction. We created interpretable visualizations, FOCUS Maps, to identify the mammogram patch that most strongly influences Mirai's risk scores. We observe that Mirai's risk estimates are primarily driven by localized imaging features and are significantly influenced by sites where cancer was subsequently detected. FOCUS Maps may assist radiologists in localizing suspicious regions for intensive imaging evaluation, such as a focused diagnostic ultrasound, bolstering its clinical utility in potentially guiding personalized screening and early intervention strategies. Although Mirai may be detecting early signs of malignancy, we also identified other localized, non-lesion imaging patterns that drive its predictions. Our explainability technique can help radiologists assess whether clinically meaningful features are associated with AI model predictions for breast cancer risk.

Overall, by developing informatics frameworks for external validation and model explainability, our work supports a comprehensive evaluation of the generalizability and trustworthiness of AI tools, potentially enabling the clinical adoption of AI tools to improve healthcare outcomes and promote health equity.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I feel sincerely grateful to have received continued support, guidance, and encouragement from my advisor, Dr. Christoph Lee. Dr. Lee, you lifted me out of my comfort zone and provided me with the opportunity to explore research directions that transformed me from a curious student into an independent research scientist. Thank you for always finding the time to speak with me and patiently clarifying all my doubts despite your extremely busy schedule as a physician-scientist. This dissertation would not have been possible without your unwavering faith in me.

I would also like to thank Dr. John Gennari, the chair of my PhD committee. Dr. Gennari, I cherish every moment I spent working with you, and I am grateful for your guidance in developing the critical thinking and writing skills that played a pivotal role in my PhD journey.

To Dr. Sean Mooney, I feel fortunate to have received your constant support. I vividly remember the first time you gave a talk to our cohort, and your advice has stayed with me ever since. Thank you for teaching me how to think at a high level and for encouraging me to pursue ambitious projects.

To Dr. Bill Lotter, it has been a privilege to work with you. I am incredibly grateful for the opportunity to pursue explainability research under your guidance. Thank you for sharing innovative ideas, encouraging me to think out of the box, and teaching me the art of scientific problem-solving. You helped me persevere when our results weren't promising. Thank you so much for expanding my research interests.

Special thanks to Dr. Kathryn Lowry for taking the time to review my dissertation and shaping my PhD aims. Brainstorming ideas with you was a wonderful learning experience. Thank you for always motivating me and patiently helping me evolve as a scientist.

I am immensely grateful to my undergraduate advisors, Dr. Chirag N. Paunwala and

Dr. Mita C. Paunwala, for introducing me to scientific research, honing my skills, and motivating me to pursue a PhD. I also thank Dr. Manpreet Katari and Dr. Dennis Shasha for taking me under your wing during my brief time at New York University. Your scientific thinking inspires me. Without your collective support, I would not have had the opportunity to pursue a PhD at the University of Washington.

I am fortunate to have received continued support and mentorship from Mr. Hemal Jani. Hemal uncle, I cannot express how grateful I am for your constant support and belief in me, which helped me succeed against all odds.

To my parents, Mona Ramwala and Ankur Ramwala, who inspired me to dream beyond boundaries. Mumma–Dadda, everything I am today is a result of your countless sacrifices. To my grandparents, whose blessings continue to be my greatest strength. To my wife, Shweta, who keeps me calm and composed and patiently endures the challenges of long distance. To my uncle, Apurv Paunwala, my aunt, Kruti Paunwala, my parents-in-law, and my cousins—Shrey, Moksha, Deeti, and Dishi—who made even the most challenging times joyous.

I would also like to thank all faculty members of the Department of Biomedical Informatics and Medical Education, and the Breast Imaging section of the Department of Radiology for insightful conversations. I am sincerely grateful to have received continued support from the NW-SCORE team, especially, Dr. Janie Lee, Suzanne Kolb, Mary Grace Asirot, and Yingke Xu. I am profoundly grateful to Savannah Partridge for introducing me to Dr. Christoph Lee. Also, many thanks to my friends from the Fred Hutch Cancer Center office, particularly, Wesley, Yui, Anum, Debosmita, and Lolly for all the thought-provoking conversations. I would also like to thank my friends from BIME—Jimmy Phuong, Peter Ju, Raghav Madan, Bhargav Vemuri, Zina Xu, Yile Chen, Michael Zhang, Chak Charoensilpchai, Amanda Tsai, Amber Chen, Ashmitha Rajendran, Brian Chang—for their camaraderie throughout this journey.

I would also like to thank, Matthew Unrath, for equipping me with skills in Amazon Web Services (AWS). Special thanks to Dr. Clara Amorosi and Dr. Adam Bleckert, my

internship managers from Bristol Myers Squibb and Vertex Pharamceuticals, respectively, for guiding me on projects with real-world impact.

My heartfelt gratitude to Neelam Dighe, Yatindra Thorat, Kailash Singh, Nihar Thakkar, Smeet Dhakecha, Ankit Bhimani, Dhiren Vyas, Ambika Patel, Yash Patel, Palak Ahuja, Himani Vachhani, and all of my friends in Seattle, who became my family. Without your support, I would not have been able to endure the 12 weeks of bed rest following multiple fractures in my leg. You helped me get back on my feet, quite literally.

I am sincerely grateful to everyone who played a pivotal role during my PhD, and I am looking forward to continued support and guidance!

DEDICATION

To my grandmother, Malini Nitinchandra Paunwala, the strongest person I know, whose breast cancer journey inspires me to never give up.

Chapter 1
INTRODUCTION

Deep learning for mammography-based breast cancer screening has been a pioneering application of AI in medical imaging [1]. However, despite encouraging early performance reports, the clinical adoption of AI tools for mammography and medical imaging has been impeded by two key challenges.

First, deep learning algorithms may underperform in new clinical settings. Hence, health care institutions should be able to validate the performance of these models on their own patient population before choosing to incorporate these models into their clinical practice. However, there is limited guidance on establishing an external validation infrastructure to compare and assess the generalizability of deep learning algorithms.

Second, adoption of AI tools hinges on the ability to understand and explain their underlying decision-making process. Unfortunately, most deep learning algorithms are black-box systems [2], making it challenging for radiologists to trust their predictions.

Our work aims to improve the clinical adoption of deep learning algorithms for mammography-based breast cancer screening by addressing these two important challenges. In this chapter, we briefly introduce the role of AI algorithms in mammography interpretation, further explain the key barriers to their clinical adoption, and summarize our original contributions to the field.

1.1 Artificial Intelligence for Mammography-based Breast Cancer Screening

Multiple randomized controlled trials have demonstrated that screening mammography decreases breast cancer mortality [3–6]. Despite the reduction in breast cancer mortality due to advances in screening and treatment [7], breast cancer remains the second leading cause of cancer-related deaths among US women [8]. The interpretive accuracy of breast cancer screening is limited by human visual perception of mammographic abnormalities, leading to

1 in 8 breast cancers being missed at the time of interpretation [9]. The high rate of missed cancers has motivated the development of methods to assist radiologists in mammography interpretation.

Traditional computer-aided detection (CAD) systems for mammography were based on hand-engineered features [10]. Based on small reader studies, CAD tools achieved clearance from the U.S. Food and Drug Administration (FDA) in 1998 and third-party payor reimbursement shortly thereafter [11]. As a result, traditional CAD was quickly adopted into routine radiologist workflows with the promise of improved performance and higher reimbursements for imaging practices [12]. Unfortunately, large-scale population studies over the next two decades demonstrated no additional benefit from using CAD for screening mammography interpretation [13, 14]. Specifically, radiologists had higher diagnostic accuracy when they did not rely on CAD for mammography interpretation.

More recently, over the last decade, the burgeoning volume of open-source mammogram imaging datasets [15–18], advances in computing power, and breakthroughs in deep learning for computer vision [19, 20] have fueled the development of artificial intelligence (AI) algorithms for mammography interpretation. In contrast to traditional feature extraction and analysis methods, deep learning algorithms employ multi-layer neural networks that automatically learn hierarchical representations of input data and model complex relationships between inputs and outputs [21]. Deep learning algorithms have achieved commendable results in various applications of mammography interpretation, including breast cancer detection [22–24] and future breast cancer risk prediction [25–27].

Beyond identifying subtle cancers present on mammography images, deep learning models also show promise in predicting future breast cancer development. Several deep learning models have demonstrated promising results in estimating multi-year future breast cancer risk directly from screening mammograms and have outperformed traditional clinical risk assessment tools, including Tyrer-Cuzick [28] and Breast Cancer Surveillance Consortium risk calculators [29]. However, traditional clinical risk assessment tools have only modest 5-year predictive accuracy for breast cancer risk, with Areas Under the Receiver Operating Characteristic (AUC) curve typically ranging from 0.60 to 0.65 [28, 30]. Newer deep learning models that make predictions solely from mammography images have demonstrated higher

predictive accuracy than traditional risk factor-based models, with some studies reporting 5-year AUCs of 0.70-0.75 [26]. If AI models with outputs based solely on mammography images are shown to have higher accuracy than traditional clinical risk tools, then AI-driven risk prediction at the time of mammography screening has the potential to improve personalized screening approaches by identifying women who would benefit from personalized prevention and screening, including chemoprevention (e.g., tamoxifen therapy) and/or supplemental screening (e.g., adding supplemental screening MRI to annual screening mammography).

1.2 Motivation and Objectives

Many AI algorithms have demonstrated promising performance for improved mammography interpretation in internal validation studies where test sets could be from the same population distribution as the training sets [27, 31–33]. However, when these models are applied to new clinical settings, they are often found to have limited generalizability [34, 35], primarily due to differences in population diversity. This poses risks for patient safety and suboptimal outcomes, particularly for patient groups that are underrepresented in the training data [36–38]. Therefore, healthcare institutions need methods to validate deep learning models on their specific patient populations before adopting them. Several challenges limit external validation efforts. First, due to patient privacy concerns, health institutions cannot share mammography imaging data with commercial AI vendors. Similarly, AI vendors may be hesitant to provide their proprietary algorithms for model validation before purchase. Second, health institutions may wish to compare multiple AI algorithms, which may have different memory and computational resource requirements. Finally, outsourcing external validation efforts to third-party institutions can result in substantial financial and legal burdens for healthcare institutions. Thus, there is a need for an open-source informatics framework to equip healthcare institutions to perform large-scale external validation of image-based AI algorithms. Ideally, such an infrastructure could be used to assess model generalizability, investigate latent biases, and compare the performance of different AI algorithms.

Beyond demonstrating generalizable performance, AI tools for mammography are less

likely to be broadly adopted if insights into their decision-making process are not available to clinicians. Most deep learning algorithms are 'black-box' systems, making it challenging to interpret their predictions [39]. Developing explainability techniques for mammography-based deep learning algorithms is even more critical for future breast cancer risk prediction because the outcome (future cancer) is not observable on the current images. Thus, there is a need for techniques that identify mammography imaging features that contribute to deep learning algorithm predictions.

To address this need, we developed a quantitative and qualitative framework to explain the predictions of Mirai [40], an open-source imaging-based deep learning algorithm, used to predict a woman's breast cancer risk over five years. Mirai has demonstrated accurate and generalizable performance in multiple external validation studies [41–44]. Our framework helps determine whether localized features, such as early signs of a developing cancer, or global features like breast density, texture, or architectural patterns, most influence Mirai's risk assessments. Explaining these predictions could lead to more effective clinical implementation of AI risk models such as Mirai by guiding radiologists to previously unrecognized signs of breast cancer risk and development.

Overall, by introducing techniques for rigorous external validation and for providing clinically meaningful explanations of AI models, our work can provide tools for health care organizations to comprehensively evaluate these promising new technologies before their integration into patient care.

1.3 Key Takeaways and Impact

To address the aforementioned needs for better approaches for mammography-based AI algorithm evaluation, we developed the following informatics methods:

1. An open-source informatics framework to enable health systems to establish a robust cloud-based infrastructure for the external clinical validation of deep learning algorithms.

We developed a framework, ClinValAI [43, 45], that meets the multiple needs of both health care organizations and algorithm developers. First, ClinValAI-based cloud infras-

tructure addresses concerns regarding patient privacy and information security, which often deter healthcare institutions from conducting external validation studies. Similarly, it allows algorithm developers to control access to their intellectual property after sharing proprietary AI models for external validation studies. To facilitate the validation of multiple AI models, ClinValAI supports scalable, customizable workflows tailored to the varying memory and computing requirements of different algorithms. Thus, ClinValAI promotes external validation efforts to assess the generalizability of AI tools before their clinical adoption.

2. A technique to provide insights into the predictions of the Mirai [26] algorithm for mammography-based future breast cancer risk prediction.

We established a method, FOCUS, to identify regions of interest (ROIs) in mammography images that drive Mirai's predictions and to determine whether these regions are localized or global. To make these findings more accessible to radiologists, we developed interpretable visualizations, FOCUS Maps, that highlight imaging regions most important to Mirai's risk assessments, thereby enabling assessment of the trustworthiness of the model's predictions. Finally, we investigated the clinical relevance of these ROIs by presenting FOCUS Maps to subspecialized breast radiologists for qualitative interpretation.

Overall, our work has the potential to improve methods for evaluating deep learning algorithms for breast cancer screening to ensure their appropriate, evidence-based adoption.

1.4 Outline

Each chapter of this dissertation addresses how we can enhance the clinical adoption of deep learning algorithms for mammography-based breast cancer screening and future risk prediction.

In **Chapter 2**, we explain our work on addressing the limited guidance for developing methods to validate the performance of AI in mammography. The material in this chapter is partially based on two publications: 'Establishing a Validation Infrastructure for Imaging-Based Artificial Intelligence Algorithms Before Clinical Implementation' by Ojas A. Ramwala, Kathryn P. Lowry, Nathan M. Cross, William Hsu, Christopher C. Austin, Sean D. Mooney, and Christoph I. Lee. in the Journal of the American College of Radiology and

'ClinValAI: A framework for developing Cloud-based infrastructures for the External Clinical Validation of AI in Medical Imaging' by Ojas A. Ramwala, Kathryn P. Lowry, Daniel S. Hippe, Matthew P. N. Unrath, Matthew J. Nyflot, Sean D. Mooney, and Christoph I. Lee, in the proceedings of the Pacific Symposium on Biocomputing. We demonstrate ClinValAI's potential by establishing a centralized cloud-based infrastructure to perform a large-scale comparative evaluation of three FDA-approved commercial AI models on screening mammograms from seven U.S. regional mammography registries affiliated with the Breast Cancer Surveillance Consortium (BCSC) [29]. Finally, we conclude by highlighting the clinical impact of our evaluation framework.

In **Chapter 3**, we advance the paradigm of developing explainability techniques faithful to the model architecture of AI algorithms. With the exception of the first paragraph, this chapter is exactly a manuscript titled 'Feature-space Occlusion to Explain a Deep Learning Breast Cancer Risk Prediction Model', currently under review at Radiology: Artificial Intelligence journal, with authors Ojas A. Ramwala, Cody Schopf, Kathryn P. Lowry, John H. Gennari, Sean D. Mooney, Christoph I. Lee, and William Lotter. We present our work on developing FOCUS (Feature-space OCclusion for Understanding Saliency), a technique that provides insights into the underlying decision-making process of Mirai [26]. We apply FOCUS to assess the importance of specific mammographic regions in Mirai's predictions. We analyzed whether its predictions are localized to specific patches of a mammogram view and whether the most representative features correlate with regions where cancer subsequently developed. We discuss the range of strategies Mirai uses to estimate future breast cancer risk and conclude by explaining the clinical impact of our work in potentially assisting breast imaging radiologists in providing more personalized screening strategies and patient-specific targeted interventions.

Finally, in **Chapter 4**, we explain—from a broader biomedical informatics and clinical perspective—the overarching contributions of our work. We specifically highlight the impact of our research in providing tools for health care organizations to rigorously evaluate AI technologies prior to clinical adoption. We acknowledge the limitations of our methods and also provide avenues for future work.

Chapter 2

EXTERNAL VALIDATION OF DEEP LEARNING ALGORITHMS FOR MEDICAL IMAGING

2.1 Motivation

Artificial Intelligence (AI) algorithms have demonstrated promising results in biomedical image [46–54] processing, yielding improved diagnostic outcomes, early intervention strategies, and well-tailored patient-specific management options. Deep learning models have demonstrated tremendous capabilities to affect radiologists' workflows [33]. AI has been transformational for several applications in radiology, with some algorithms performing comparably or even superior to radiologists on specific tasks [31]. AI algorithms for breast cancer detection on screening mammography, for instance, have outperformed radiologists in interpreting mammograms in controlled study environments [32]. Combining the outputs of AI models with radiologist interpretation has also led to substantial performance enhancements in breast cancer detection [27, 33]. As of December 2025, 1,357 AI algorithms have received FDA clearance, with 1,039 for applications in Radiology [55]. These algorithms are cleared for use in clinical settings.

Nevertheless, AI models can experience reduced performance in new clinical settings and pose challenges to achieving the desired population-level outcomes [34]. The limited generalizability of deep learning algorithms can raise concerns about patients' safety [36]. The performance of AI models can vary across diverse subpopulations [37], leading to disparities in health outcomes [38]. Research studies that independently validate the performance of AI algorithms on imaging data representative of real-world settings have reported performance reduction for certain subpopulations [35]. Biased training datasets used for developing AI models can exacerbate the fairness of algorithms when applied to new populations. The adoption of such algorithms can have critical implications for patients' safety.

Thus, developing and leveraging infrastructures that can effectively validate the perfor-

mance of AI algorithms on specific target populations is crucial to enhancing the potential of integrating AI algorithms into radiology clinical workflows. However, rigorous external validation of AI algorithms is impeded by numerous challenges at any given institution. Algorithm vendors may be hesitant to share commercial AI models with medical centers for independent validation before purchasing the tool. Moreover, per HIPAA guidelines, academic institutions cannot provide imaging data to AI vendors to safeguard patient privacy and avoid ethical concerns about the potential use of clinical data to improve commercial algorithms. Moreover, different AI algorithms have varying storage and computing requirements. Planning and budgeting for resources to cater to such varying needs can cause substantial financial and cognitive burdens on health systems evaluating multiple AI tools on-premises for clinical adoption. Although health institutions can potentially outsource the work of validating AI algorithms, procuring outside services can be even more expensive and can entail cumbersome legal and technical challenges for providing access to patient imaging data.

Our work in this chapter equips large practices and health care institutions with a framework for establishing robust infrastructure to support rigorous, comparative external validation of multiple potential AI algorithms. This framework supports evaluating these algorithms for specific clinical scenarios before purchase and/or clinical adoption. The overall goal of this external validation framework is to enhance the decision-making process for identifying the most suitable AI model for specific applications locally, ensuring that the optimal AI algorithm is chosen to improve the health of the target population.

2.2 Guidelines for establishing infrastructures for the external validation of AI algorithms for medical imaging

To develop efficient, customizable, and cost-effective infrastructures for the external validation of AI models, it is essential to understand the associated challenges and provide guidelines on addressing them. In this section, we summarize the guidelines that we published in [45]. These guidelines help healthcare institutions establish an AI external validation infrastructure to assess the local performance of AI algorithms before health practice or system-wide implementation.

2.2.1 On-Premises versus Cloud-based Infrastructures

Health care organizations can validate the performance of AI algorithms either by installing the models on site or by hosting them on a cloud infrastructure. To achieve enhanced protection of patient data, health care organizations may be particularly inclined to set up on-premises solutions to leverage their more secure and tangible nature. Likewise, IT teams may be concerned about delays in response times and may prefer on-premises infrastructures for their lower latencies and real-time processing capabilities. However, the limited scalability of on-premises infrastructures poses challenges when evaluating multiple AI algorithms. Adapting an on-premises setup to provision for the diverse computational specifications of multiple models may require procuring additional hardware resources and may entail configuring them to be compatible with the existing clinical setup. Ancillary data backup solutions will need to be established to address the susceptibility of on-premises setups to hardware failures and data loss. In addition to potential technical difficulties, the increased expenses, including the up-front capital investment in hardware equipment, can pose significant financial barriers. The alternative to an on-premises setup is developing cloud-based solutions to use their flexibility in customizing computing and storage resources on the basis of the diverse requirements of various AI models. There are no up-front large expenditures when setting up a cloud-based infrastructure [56]. The ability to scale down resources after executing specific validation steps further improves its potential cost-effectiveness. Institutions can take advantage of reduced maintenance processes because cloud services have built-in resource management solutions that are regularly updated to circumvent the requirement of manual interventions. Cloud setups facilitate robust data backup and recovery solutions to safeguard against inadvertent information loss. Furthermore, cloud solutions enable more seamless deployment of AI algorithms for evaluating them on large-scale datasets. Additional advantages, such as quick procurement, minimal on-site energy consumption, and remote access, make cloud setups an appealing choice. However, it is crucial to address the risks associated with uploading patient data to the cloud.

2.2.2 Patient Privacy Concerns

Infrastructures for rigorous external validation of AI algorithms must be equipped to address security concerns associated with protected health information [57]. Because permitting algorithm vendors access to an institution's clinical data infrastructure behind a firewall may constitute a data security risk, a model-to-data [58, 59] framework should be established. By leveraging this approach, in which AI models are shared with institutions rather than patient imaging data being shared with vendors, the risks of exposing protected health information can be circumvented or significantly reduced. Medical imaging data should remain securely within the health care organization's infrastructure with controlled access. AI vendors can provide their algorithms with all necessary dependencies and packages via a Docker image for AI algorithm evaluation. Moreover, the algorithms can be supplemented with a license file, allowing vendors to manage access to their proprietary systems.

2.2.3 Data Collection and Curation

Appropriate collection of high-quality imaging data is pivotal to ensuring a faithful pipeline for validating AI models. The data distribution must reflect the real-world target population, and the statistical plan and sampling techniques should account for an adequate sample size to look at AI performance in subpopulations. Subpopulations of concern at a given institution must be identified prospectively to assess prevalence in the general population to ensure an adequate sample size. By addressing the diversity of the imaging data up front, the potential of enhancing equitable health outcomes is increased.

After finalizing the data for the external validation analysis, secured mechanisms for transferring de-identified imaging data from PACS servers to cloud-based storage should be established. A comprehensive metadata file should be drafted to verify that all imaging data has been uploaded to storage. Essentially, a spreadsheet comprising de-identified examination numbers can be used to match the transferred batch of images and eventual AI outputs. This manifest file can also aid in organizing and aggregating data before statistical analysis.

2.2.4 Understanding prerequisites of AI algorithms

Rigorous external validation entails working with various algorithm-specific requirements. Health institutions aiming to identify appropriate AI models for their particular application need to establish a standardized protocol to fulfill those prerequisites. Because deep learning models have different parametric complexities, their computational requirements can vary. Receiving thorough documentation from algorithm vendors can facilitate the process of establishing necessary computational resources. Computing instances in the cloud environment can be effectively mapped to the resource requirements and can be temporarily acquired and customized. Similarly, diverse AI algorithms can employ varying image processing mechanisms and have different memory requirements. Thus, computational resources based on the documentation received from the vendors must be established to implement external validation infrastructures for AI.

In addition to the pixel array information stored in the DICOM metadata, other tags may be important to the algorithmic scoring of imaging examinations. For example, AI algorithms for breast cancer detection on mammograms may necessitate DICOM tags for image laterality (left or right breast), manufacturer details (GE, Hologic, etc), and image projection information (craniocaudal or mediolateral oblique) to accurately provide an examination-level AI score for malignancy risk. Because AI models can necessitate the inclusion of disparate DICOM metadata information for their respective computational workflows, vendor documentation should also include an exhaustive list of DICOM headers necessary for executing their scoring pipelines. If deidentified images are used for validating AI algorithms, it is imperative to check for accidental deletion of the necessary DICOM headers essential to the scoring pipeline. Thus, a dedicated review mechanism, incorporating input auditing logic, should be established to automatically scan DICOM metadata and detect missing tags, thereby supplementing the images with necessary information. Overall, the auditing logic should ensure that the imaging data are in the appropriate format for algorithmic processing.

AI algorithms can ingest imaging data via different file structures. Similarly, radiology departments can have distinct data structuring mechanisms in their respective PACS

servers. Establishing coherence between the organization of the imaging data and the structuring requirements of AI models is imperative for accurate scoring. External validation infrastructures should be customized to represent the various modes of organizing DICOM files and must be equipped to rearrange images according to the algorithm's prerequisites. For example, suppose each image is organized into individual folders in the PACS servers of a health institution. In that case, the external validation infrastructure should restructure the DICOM files based on imaging modality, date of acquisition, patient ID, and other parameters, depending on the specific AI model's requirements. The documentation received from the AI vendors should include precise information on image structuring.

2.2.5 Comparative Scoring of AI Algorithms

To reliably validate AI models, comprehensive documentation explaining the implementation steps necessary for an AI algorithm to process each medical imaging exam must be received. The computational environment of the cloud computing environment should use Docker images of the AI models, so that they are ensured to include all necessary libraries to run the AI models on the imaging data. The imaging data can be mounted onto the corresponding Docker container, and the necessary computational programs, per the documentation, can be sequentially executed to validate AI algorithms on imaging data.

After processing a medical imaging exam, AI models can generate a plethora of output files and results. Vendor documentation can outline the meaning of these files and must detail the exact steps to be followed to receive the final scores. Because various AI models can express their outputs through different representations—for example, comma-separated values files, JavaScript Object Notation objects, or DICOM structured reports—the infrastructure must be equipped with scripts to extract numerical values of interest from the outputs. A comprehensive explanation corresponding to the numeric scores must also be received to aid the radiologists in understanding the AI algorithm's interpretation of the imaging data. Moreover, the AI model can also generate secondary outputs, for example, heatmaps, annotated images, bounding box coordinates, and others. The external validation infrastructure should also ease the process of capturing these supplementary files so that they can be referenced as needed to increase the radiologists' confidence in a particular

algorithm's reasoning.

The infrastructure can be equipped with parallel scoring workflows to efficiently leverage computing power. Essentially, depending on the availability of computational resources, the entire dataset can be split into multiple batches, with every batch being allocated a dedicated computing instance. The infrastructure can be configured to reorganize, score, and retrieve final outputs for all examinations in a batch in a serial fashion. A dedicated mechanism to integrate serial and parallel pipelines can enhance the overall validation process. These batch-processing workflows can be especially valuable and cost-effective while using cloud-based infrastructures.

After the completion of the scoring processes, the external validation infrastructure must confirm that scores have been generated for all examinations. The infrastructure should be equipped to capture algorithmic error messages and interpret infrastructure logs to iteratively update the scoring mechanism as appropriate and rerun the algorithms on examinations that generate errors. The infrastructure should also generate a list of output files that could have been corrupted because of anomalies in the cloud computing environment to distinguish them from invalid examinations. In addition to automated output verification, manual checks may be necessary to establish the validity of the external validation infrastructure. These can be implemented through an output verification logic in the external validation infrastructure. To thoroughly evaluate the scoring mechanism, some example imaging data can be received from the AI vendors, if available, to confirm that the infrastructure outputs match the expected scores. Moreover, the statistical distribution of the final outputs should be comprehensively analyzed for the detection of outliers. These outliers should undergo troubleshooting to identify additional corrections required in the external validation infrastructure.

Downstream statistical evaluation can be performed to record performance metrics, including diagnostic accuracy measures such as the area under the receiver operating characteristic curve or the F1 score. More clinically meaningful outcome measures may include outcomes such as true-positive rates, false-positive rates, and false-negative rates. These outcome measures will be application-specific and should be compared in a pre- versus post-AI intervention analysis [33, 60].

In summary, establishing dedicated mechanisms for safeguarding patient privacy, obtaining vendor cooperation for evaluation before purchasing new AI tools, managing imaging data collection, resource allocation, and algorithm inference are all major tasks required for developing infrastructures to comprehensively evaluate AI model performance. Thus, these guidelines promote an evidence-based approach for adopting AI models that can enhance radiology workflows and improve patient outcomes. As we describe next, we leveraged these guidelines when developing ClinValAI. We demonstrate its utility by performing external clinical validation of Mirai, a mammography-based breast cancer risk prediction algorithm, and a large-scale comparative external validation of three FDA-approved mammography-based breast cancer detection algorithms.

2.3 ClinValAI: A framework for developing Cloud-based infrastructures for the External Clinical Validation of AI in Medical Imaging

To address the challenges described above, we developed ClinValAI30 – an open-source cloud-agnostic unified framework for establishing robust infrastructures to validate AI algorithms. We customize its functionalities for the clinical validation of AI models for medical imaging applications. Using ClinValAI, medical institutions can rigorously evaluate models before integrating them into clinical workflows. Healthcare institutions can use our framework to screen data from large populations to accurately assess model generalizability and investigate latent biases.

ClinValAI can be used to establish innovative, effective, and secure cloud-based validation infrastructures. Figure 2.1 details our conceptual framework for externally validating AI models in clinical medicine.

2.3.1 Preserving Patient Data Privacy

Patient privacy and information security concerns constrain biomedical data sharing and stymie AI algorithm development and validation efforts [61]. ClinValAI leverages the "Model to Data" (MTD) paradigm [58, 62] to validate AI models on private biomedical data. Cloud infrastructure and containerization techniques form the foundation of the MTD framework. Rather than providing direct data access to the vendors, the Dockerized models are uploaded

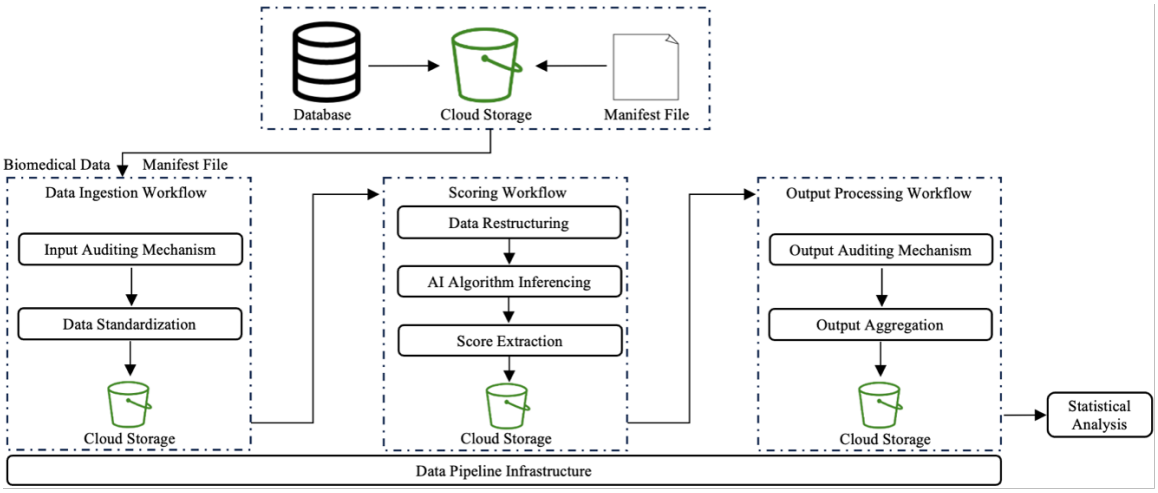


Figure 2.1: Conceptual overview of the ClinValAI framework.

to the cloud host as containers encapsulating the AI algorithms, their dependencies, and other configuration settings required for successfully testing the models on the data stored in the cloud. To address intellectual property concerns, ClinValAI supports license files for Docker images, allowing AI vendors to control access to their AI models. Thus, ClinValAI enables health institutions to preserve patient data within a firewall and run models on medical imaging exams without providing vendors direct data access.

2.3.2 Data Pipeline Infrastructure

ClinValAI features multiple computational pipelines for biomedical data processing and clinical validation of AI algorithms through a combination of series and parallel jobs.

Workflow Representation

To comprehensively express the workflow design, we leverage the Workflow Description Language (WDL) [63] due to its comprehensibility and cross-platform interoperability. WDL enables defining pipelines to process and analyze data. WDL necessitates an engine to execute its functionalities. Our proposed framework utilizes miniWDL [64], a WDL execution engine for biomedical applications that functions as a job orchestrator for executing multiple data processing workflows in a parallel fashion, depending on the available memory and computing resources. The customizability of ClinValAI's workflow representation method

bolsters its utilization for the clinical validation of AI.

Job Scheduling and Batch Processing Orchestration Mechanism

Our framework is equipped with tools that provision compute instances and communicate with the miniWDL engine and a container job scheduling mechanism to automate infrastructure deployment (Figure 2.2). It can be further modified for more granular control over those pipelines, as described in Figure 2.1. After the workflow submission, the WDL script is uploaded to cloud storage, and the job scheduling mechanism is invoked to run a miniWDL container, known as the "head" job container. ClinValAI implements data processing pipelines through the miniWDL engine operating on this container. The head job pulls the WDL script from the cloud storage and, per its instructions, directs the scheduling mechanism to spin up "task" job Docker containers that execute individual components of the workflow. ClinValAI enables the head job containers to spin up multiple sets of task job containers to achieve the parallel execution of computational steps.

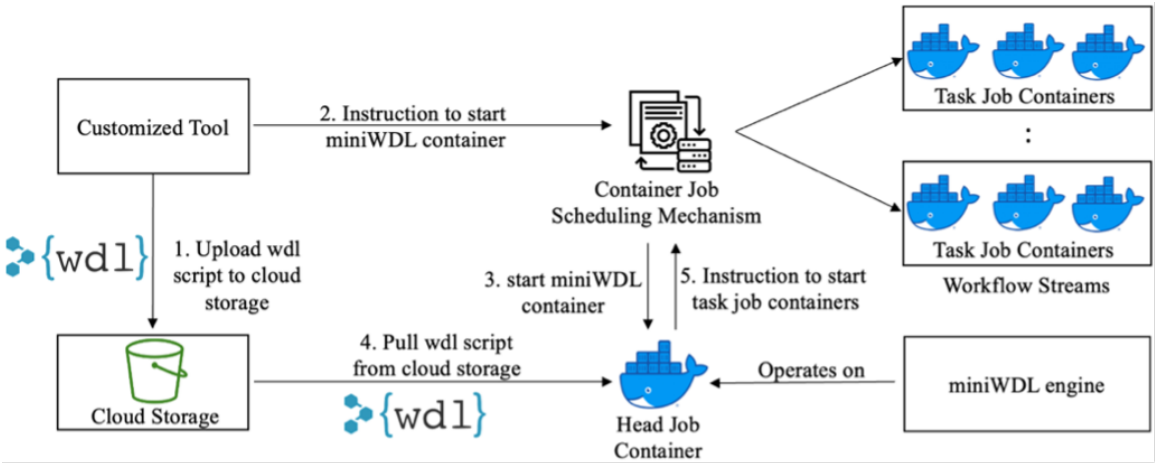


Figure 2.2: ClinValAI's job scheduling and batch processing orchestration mechanism.

For the external validation of mammography-based breast cancer screening algorithms, our ClinValAI-based cloud infrastructure ingests a set of compressed files, each representing a batch comprising multiple sub-folders corresponding to patients' mammography exams. ClinValAI creates multiple execution streams for each set; all exams in one batch are processed serially by leveraging numerous task containers running sequentially. Exams in one

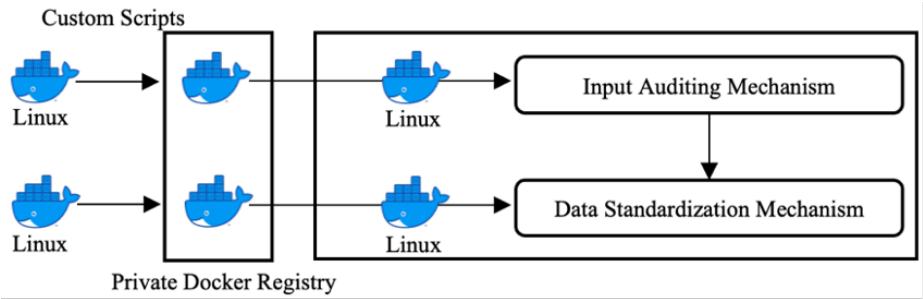


Figure 2.3: ClinValAI's data ingestion workflow ensures that inputs are consistent with model prerequisites.

batch are scored independently of other batches in a parallel fashion. Thus, ClinValAI's data pipeline enables leveraging the full potential of the cloud computing environment.

In addition to validating AI models using their Docker images, our framework supports customizing Linux Docker images to establish optimized workflows. Rather than dynamically pulling scripts from cloud storage at run-time, ClinValAI facilitates configuring the Docker images at build time. This approach avoids inadvertent version updates in the sequence of instructions during run-time, which could produce inconsistent results. While fetching scripts from cloud storage during run-time is more convenient, baking them into the Docker image enhances reliability.

2.3.3 Data Ingestion Workflow

ClinValAI's data ingestion workflow (Figure 2.3) is the first of the three stages in the framework. It comprises an input auditing and a data standardization mechanism.

Input Auditing Mechanism

ClinValAI's input auditing mechanism performs the vital task of verifying if the data can be processed and is aligned with the model's prerequisites before initiating the scoring process. This can help ensure a sample size that preserves statistical power for meaningful analysis. Through a configured Linux Docker image, it compares the uploaded data with a manifest file and algorithm-specific requirements to verify that the dataset is complete with all the required information.

For external validation studies, a manifest file is created that comprises the accession

numbers, data modality, the corresponding number of images in each exam, image laterality and projection, file sizes, and other relevant information. The auditing logic checks for corrupted files, DICOMs with missing pixel array data, and unsupported manufacturing devices, and monitors if the image metadata contains all the information required by the algorithm. For example, AI models for mammography interpretation may not be able to process images if view/projection (Cranio-Caudal (CC) or Medio-Lateral Oblique (MLO)) or laterality (left or right breast) information is missing from DICOMs. ClinValAI thoroughly analyzes the data to identify such inconsistencies and features a comprehensive input auditing mechanism to ensure a seamless external validation study.

Data Standardization

Standardizing inputs before initiating AI algorithm processing is necessary if there is variation from multiple data sources or if a data source requires enrichment before algorithmic processing can take place. ClinValAI's data standardization mechanism analyzes the findings of the input auditing logic and provides the feature of customizing the associated Linux Docker image to achieve data standardization and ensure the quality of the study data.

For external validation of AI models for mammography interpretation, if a set of DICOMs is corrupted or missing pixel array information, the standardization mechanism does not pass them through the scoring workflow. Similarly, it removes images that do not match the study criteria – for example, deleting all the non-mammography images to ensure that only the acceptable modalities are included. One of the important aspects of ClinValAI's data standardization mechanism is its ability to impute missing information. For example, if an image does not have laterality or projection information in the DICOM headers, the framework populates the DICOM metadata using the details from the manifest files.

Moreover, if the required data is not available in the manifest file, it parses other descriptive DICOM headers to look for specific information for imputation. For example, AI algorithms for mammography interpretation expect laterality information in one of the ImageLaterality, Laterality, or FrameLaterality headers and projection information in the ViewPosition header. If these tags are missing, ClinValAI's data standardization mechanism analyzes other subjective headers like SeriesDescription to systematically impute laterality

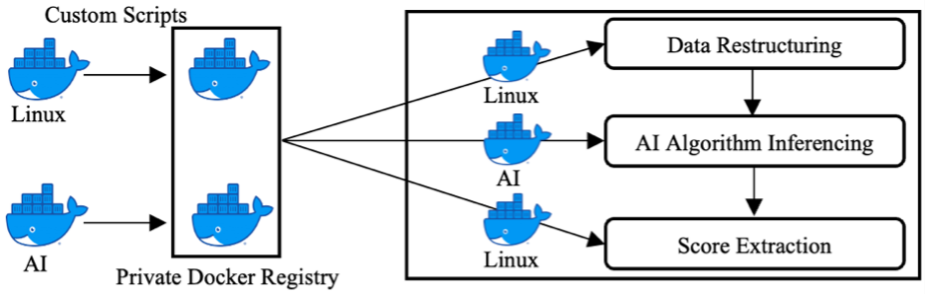


Figure 2.4: ClinValAI's scoring workflow ensures consistent inferencing and systematic troubleshooting

and projection information into their respective tags. Thus, ClinValAI can be customized to facilitate effective data cleaning and preprocessing, information imputation, and data standardization.

2.3.4 Scoring Workflow

ClinValAI's scoring workflow (Figure 2.4) is the second stage in the framework. It comprises a data restructuring, an algorithm inferencing, and a score extraction mechanism.

Data Restructuring

Various health institutions can organize patient data and medical images in different formats, requiring datasets to be systematized by patient ID, accession numbers, or date of collection. Different AI models can have their specific input structuring requirements. For example, a model may require all images from a patient to be in a single folder, while another may need additional sub-folders based on exam ID or modality. Different models may need varying numbers of images per exam – for instance, a breast cancer screening algorithm may need four standard 2D views of a mammogram (CC and MLO views of the left and right breast), whereas some models can function even with unilateral exams. Some models can raise errors if inputs contain multiple images of the same view and laterality combination, while others can successfully score them. Moreover, some models can process 2D and 3D images simultaneously, while others can leverage separate Docker images depending on shape and modality. ClinValAI supports extensive data restructuring by enabling the customization of Docker images to account for model-specific variations through the holistic

analysis of DICOM metadata and pixel array information, thereby establishing consistency between input and model criteria.

AI Algorithm Inferencing

ClinValAI enables effective customization of AI algorithms' Docker images to facilitate accurate AI algorithm inferencing, i.e., using the trained AI model to generate predictions on input data. The Docker file is specified with the required environment variables and necessary scoring scripts, and the updated Docker image is used to spin up the AI model's Docker container to execute algorithmic processing. Information about the computational requirements of the AI algorithms can be utilized to identify the appropriate compute instances to be specified in our framework. To work with asynchronous inferencing workflows, our framework also features a polling mechanism depending on the inference time of each algorithm to ensure that the compute instances are not stalled due to inconsistent data, node failures, or other issues. Furthermore, our framework provides the flexibility of incorporating additional steps, such as drafting a list of input studies to be processed or creating corresponding output folders for storing final results, depending on the models' prerequisites. Thus, by facilitating multiple customization features, ClinValAI enables robust validation of AI algorithms.

Score Extraction

After the completion of the scoring process, the model's generated files need to be processed to retrieve specific outputs of interest, such as image-, exam-, or patient-level scores. Different AI algorithms have different ways of representing outputs. ClinValAI enables customizing the Linux Docker image to follow the modes and steps to extract scores from diverse formats – from flat files like comma-separated values (CSV) documents to highly nested DICOM Structured Reports (SRs) and JavaScript Object Notation (JSON) objects. Similarly, ClinValAI also facilitates the storage of supplementary files, such as annotations in processed images or heat maps, and associated model explanations, if available, to facilitate improved interpretation for radiologists. Moreover, this step also records and organizes logs specific to the algorithm and workflow. Thus, ClinValAI facilitates systematic troubleshooting, effective scoring, and rigorous clinical validation of AI algorithms.

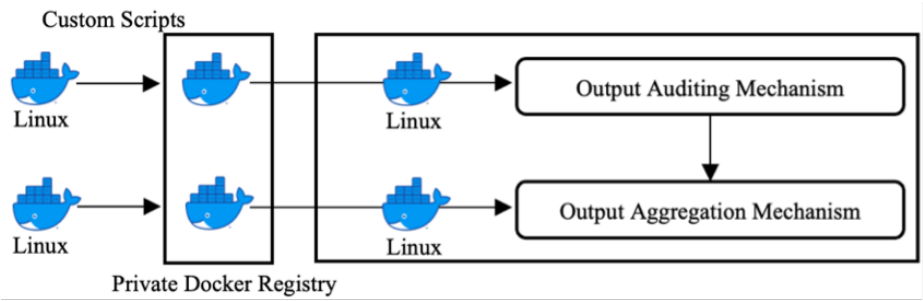


Figure 2.5: ClinValAI's output processing workflow helps identify and analyze samples with missing results and aggregates final outputs for downstream analysis.

2.3.5 Output Processing Workflow

ClinValAI's output processing workflow (Figure 2.5) is the third and final stage in the framework. It comprises an output auditing and an output aggregation mechanism.

Output Auditing Mechanism

Once the scoring workflow has been executed, ClinValAI performs the essential task of verifying if results have been produced for all the inputs and if the generated files comply with the algorithm's expected outputs. Moreover, the framework facilitates examining if the required numeric values of interest, inference reports, and supplementary files can be extracted from the resulting outputs. ClinValAI identifies samples with missing output data, irretrievable scores, and corrupted output files to enable analysis of samples to be re-scored. If no outputs are generated for a patient's exam, infrastructure-specific logs can be inspected to check for issues related to compute instances or customization of the Docker images. If scores cannot be extracted from the model's output for an exam, algorithm-specific logs can be analyzed to check for inconsistencies and errors. Overall, ClinValAI facilitates a holistic output auditing mechanism for the streamlined validation of models.

Output Aggregation Mechanism

Outputs from individual workflows are hierarchically stored based on set number, batch number, and exam ID. Analyzing the complete dataset in the distributed format of cloud storage can be cumbersome. Before statistical analysis can be performed, ClinValAI system-

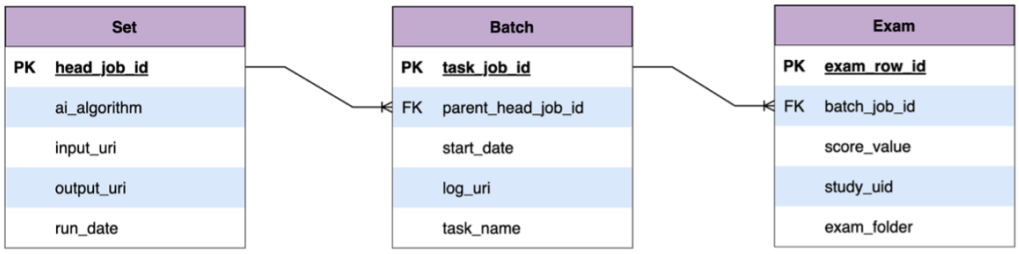


Figure 2.6: Entity Relationship Diagram for the aggregated output data. Abbreviations: URI=Uniform Resource Identifier; PK=Primary Key; FK=Foreign Key

atically aggregates relevant details by appending all results to a relational database. After the completion of scoring workflows for all standardized batches of exams, the pipeline connects to the database and hierarchically uploads data from the audited results, supplementary files, and logs by inserting rows for every set, batch, and exam as demonstrated by the entity relationship diagram (Figure 2.6). During statistical analysis, this database is pulled to analyze findings. ExamIDs and Study UIDs (Unique Identifiers) are used to cross-reference the AI algorithm's results and the attributes of interest.

Thus, ClinValAI features multiple customizable workflows to establish optimized infrastructures for robust clinical validation of AI algorithms for medical imaging applications.

2.4 Demonstrating the utility of ClinValAI

To evaluate the utility of our framework, we used our ClinValAI-based cloud infrastructure to perform a rigorous external validation of Mirai [26], a state-of-the-art open-source deep learning algorithm that predicts future breast cancer risk across five years by processing the four standard views of a 2D digital mammogram – Cranio-Caudal and Medio-Lateral Oblique views of the left and right breast. Additional details on Mirai's model architecture are explained in Chapter 3.

2.4.1 Patient Cohort

All mammography screening exams from 2010-2014 performed across four imaging facilities in the University of Washington (UW) Medicine health system were reviewed for eligibility.

Exams of women with age < 40 or ≥ 80 years, a personal history of breast cancer, or the presence of breast implants were excluded. Cancer outcomes at year 5 after every exam were collected via linkage to the Washington State cancer registry, which captures all breast cancers diagnosed within the state of Washington through December 31st, 2020, allowing for robust ground truth for all screening exams. Information on breast density and patient demographics, including age at the time of imaging and race, were obtained from the University of Washington Medicine electronic medical records. ClinValAI excluded exams with insufficient 2D screening images and processed 26,449 exams from 14,291 patients to generate Mirai scores. A total of 543 exams (2.1%) were followed by a breast cancer diagnosis within five years (88 in year 1, 92 in year 2, 112 in year 3, 119 in year 4, and 132 in year 5). Table 2.1 shows the patient characteristics. BI-RADS[65] categories 'heterogeneously dense' and 'extremely dense' correspond to dense breasts, and 'almost entirely fatty' and 'scattered fibroglandular' correspond to non-dense breasts.

2.4.2 Statistical Analysis

A mammography exam was used as the unit of analysis. Nonindependence of multiple exams from the same women was accounted for in calculations of 95% confidence intervals (CIs) and p-values by using generalized estimating equations (GEE) or the nonparametric bootstrap, clustered by woman [66]. The Mirai algorithm provides cumulative risk predictions for years 1-5 following the index examination. The outcome used for evaluating the performance of Mirai was the presence/absence of a cancer diagnosis at each timeframe. The discrimination performance of Mirai was evaluated using receiver operating characteristic (ROC) curves, the area under the ROC curve (AUC), and Uno's concordance index (c-index) as an overall summary over the 5-year timeframe [67]. The calibration of Mirai was evaluated using calibration plots and corresponding summaries of overall calibration (calibration-in-the-large) and the calibration slope [68]. To help distinguish between breast cancer detection vs. risk prediction performance, we performed the analyses using all available exams and then repeated the analyses after excluding exams that had a breast cancer diagnosis within six months. All statistical analyses were conducted using R (version 4.3, R Foundation for Statistical Computing, Vienna, Austria). All hypothesis tests were two-sided, with

Table 2.1: Patient characteristics at each exam.

		Breast Cancer within 5 years	
	All	**Yes**	**No**
Variable	**(n = 26,449)**	**(n = 543)**	**(n = 25,906)**
Age			
40-49	7,014 (26.5%)	114 (21.0%)	6,900 (26.6%)
50-59	9,431 (35.7%)	151 (27.8%)	9,280 (35.8%)
60-69	7,082 (26.8%)	171 (31.5%)	6,911 (26.7%)
70-79	2,922 (11.0%)	107 (19.7%)	2,815 (10.9%)
Race			
White	20,365 (82.6%)	460 (87.1%)	19,905 (82.5%)
Black	1,649 (6.7%)	31 (5.9%)	1,618 (6.7%)
Asian	2,394 (9.7%)	33 (6.2%)	2,361 (9.8%)
Other	241 (1.0%)	4 (0.8%)	237 (1.0%)
Unknown	1,800	15	1,785
Breast density			
Not dense	11,659 (44.1%)	216 (39.8%)	11,443 (44.2%)
Dense	14,786 (55.9%)	327 (60.2%)	14,459 (55.8%)
Unknown	4	0	4

Values are number (%).

statistical significance defined as $p < 0.05$.

2.4.3 Discrimination Performance

AUCs ranged from $0.81(95\%CI : 0.75 - 0.86)$ for 1-year cancer outcomes with the 1-year Mirai scores to $0.70(95\%CI : 0.0.67 - 0.72)$ for 5-year cancer outcomes with the 5-year Mirai scores when including all examinations (Figure 2.7, Table 2.2). The c-index was $0.70(95\%CI : 0.67 - 0.72)$. After excluding 70 exams with a cancer diagnosis within six months, the AUC was $0.72(95\%CI : 0.56 - 0.84)$ at 1 year and $0.68(95\%CI : 0.65 - 71)$ at 5 years, while the c-index was $0.68(95\%CI : 0.65 - 0.70)$. These values were more similar to previously reported results in other cohorts [26, 41] after applying the same type of exclusion [41] (Table 2.2), though they were still on the lower end of the range. Discrimination, as

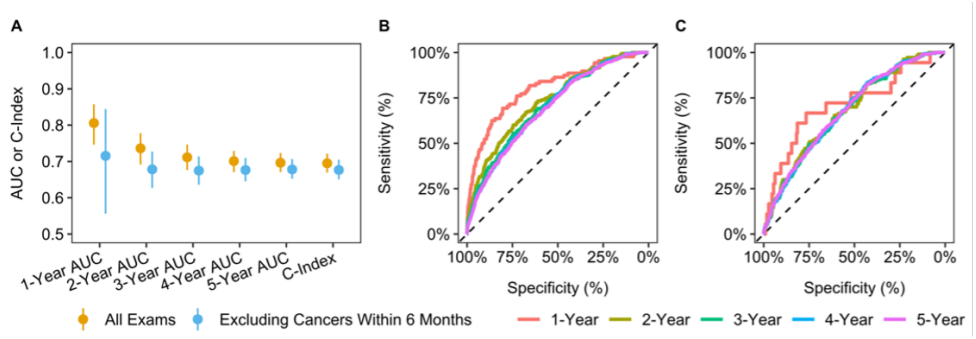


Figure 2.7: Discrimination performance of Mirai. Panel A: ROCAUC values over time and the overall c-index. The orange values are based on all exams, and the blue values are after excluding cancers within six months. Error bars represent 95% CIs. Panel B: ROC curves at different time points based on all exams. Panel C: ROC curves at different time points after excluding cancers within six months.

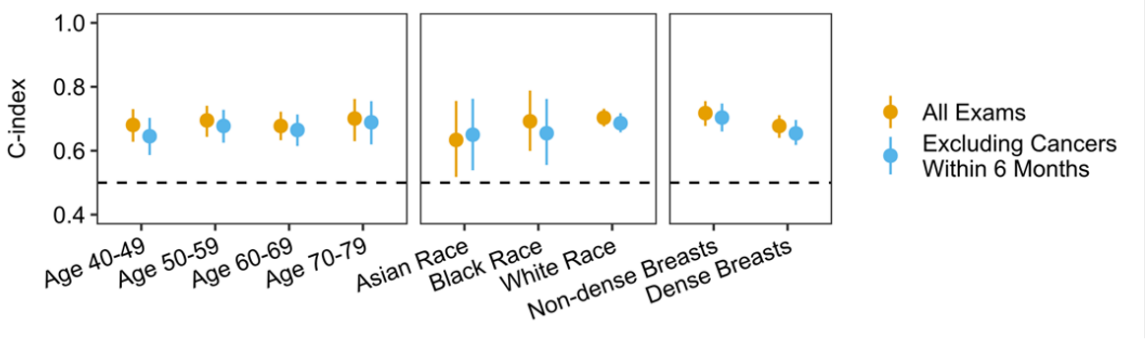


Figure 2.8: Discrimination performance of Mirai within subgroups. Error bars: 95% CIs. Dashed line: AUC = 0.50.

measured by the overall c-index, was also examined within subgroups defined by age, race, and breast density, as shown in Figure 2.8. There were no statistically significant differences in the c-index between subgroups (unadjusted $p > 0.094$ for each comparison).

2.4.4 Calibration Performance

Calibration plots for Mirai risk predictions versus observed at different timeframes are shown in Figure 2.9. The corresponding metrics of overall calibration (observed risk minus mean

Table 2.2: Discrimination performance of Mirai in the University of Washington and 7 previously reported cohorts.

	1-Year AUC (95% CI)	5-Year AUC (95% CI)	C-index (95% CI)
All Exams			
University of Washington, USA	**0.81 (0.75–0.86)**	**0.70 (0.67–0.72)**	**0.70 (0.67–0.72)**
MGH, USA[26]	0.84 (0.80–0.87)	0.76 (0.73–0.79)	0.75 (0.72–0.78)
Novant, USA[41]	0.78 (0.73–0.84)	0.75 (0.70–0.80)	0.75 (0.70–0.80)
Emory, USA[41]	0.83 (0.81–0.86)	0.76 (0.74–0.79)	0.77 (0.75–0.79)
Maccabi-Assuta, Israel[41]	0.86 (0.81–0.91)	0.75 (0.71–0.79)	0.77 (0.73–0.81)
Karolinska, Sweden[26]	0.90 (0.89–0.92)	0.78 (0.76–0.80)	0.81 (0.79–0.82)
CGMH, Taiwan[26]	0.90 (0.87–0.93)	0.79 (0.75–0.82)	0.79 (0.76–0.83)
Barretos, Brazil[26]	0.89 (0.86–0.93)	0.82 (0.78–0.86)	0.84 (0.81–0.88)
Excluding Cancers within 6 Months			
University of Washington, USA	**0.72 (0.56–0.84)**	**0.68 (0.65–0.71)**	**0.68 (0.65–0.70)**
MGH, USA[26]	0.71 (0.60–0.84)	0.71 (0.68–0.75)	0.69 (0.66–0.73)
Novant, USA[41]	N/A	0.72 (0.66–0.79)	0.72 (0.66–0.79)
Emory, USA[41]	0.74 (0.66–0.84)	0.71 (0.68–0.74)	0.69 (0.66–0.72)
Maccabi-Assuta, Israel[41]	N/A	0.68 (0.62–0.74)	0.70 (0.64–0.76)
Karolinska, Sweden[26]	N/A	0.71 (0.69–0.73)	0.71 (0.69–0.74)
CGMH, Taiwan[26]	0.84 (0.72–0.99)	0.70 (0.66–0.75)	0.70 (0.66–0.75)
Barretos, Brazil[41]	0.87 (0.80–0.94)	0.75 (0.70–0.80)	0.78 (0.74–0.83)

MGH = Massachusetts General Hospital; CGMH = Chang Gung Memorial Hospital.

predicted risk) and the calibration slope are shown in Table 2.3. When all exams are included, the metrics indicated significantly overestimated risk in years 1-2 (overall calibration: -0.15% to -0.10%, $p < 0.014$ for both), but that Mirai was overall reasonably well calibrated for the later years, where the 95% CIs for overall calibration included zero (no difference between observed and predicted risk on average) and the 95% CIs for the

Table 2.3: Calibration statistics for Mirai.

	All Exams			After Excluding Cancers within 6 Months		
	Overall Calibration*		Calibration Slope†	Overall Calibration*		Calibration Slope†
Timeframe	Estimate (%)	(95% CI)	Estimate (95% CI)	Estimate (%)	(95% CI)	Estimate (95% CI)
1-year risk	−0.10	(−0.17, −0.03)	0.83 (0.52, 1.22)	−0.36	(−0.39, −0.32)	0.05 (0.00, 0.12)
2-year risk	−0.15	(−0.25, −0.04)	0.76 (0.51, 1.06)	−0.40	(−0.48, −0.32)	0.21 (0.08, 0.36)
3-year risk	−0.13	(−0.29, 0.03)	0.89 (0.60, 1.20)	−0.38	(−0.53, −0.26)	0.40 (0.20, 0.63)
4-year risk	−0.05	(−0.25, 0.14)	0.90 (0.63, 1.20)	−0.31	(−0.49, −0.12)	0.50 (0.27, 0.76)
5-year risk	0.09	(−0.15, 0.33)	1.03 (0.75, 1.32)	−0.16	(−0.38, 0.07)	0.66 (0.44, 0.91)

*Observed risk minus mean predicted risk; a value > 0 indicates the prediction underestimates risk on average, and a value < 0 indicates the prediction overestimates risk.

†A well-calibrated model has a calibration slope of 1; slope > 1 indicates that high predictions tended to underestimate risk (not high enough) and low predictions tended to overestimate risk (not low enough); slope < 1 indicates predictions tended to be more extreme than observed (high values too high and low values too low).

calibration slope included 1 (predictions were not more or less extreme [farther from the mean] than observed on average).

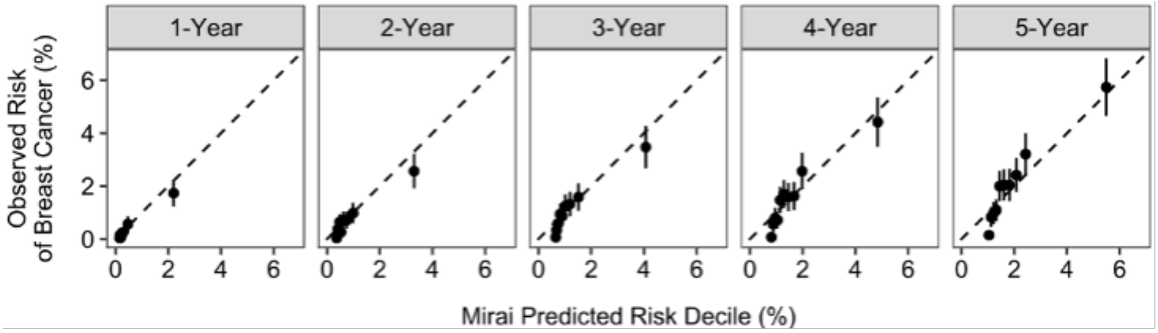


Figure 2.9: Calibration plots of Mirai with all exams included, with predicted risks grouped into deciles of approximately equal size. Error bars represent 95% CIs. The dashed line corresponds to perfect calibration (intercept = 0, slope = 1).

When exams with cancer diagnoses within six months were excluded, the calibration metrics substantially worsened (Table 3). Overall, Mirai significantly overestimated risk, more so at earlier timeframes (overall calibration: -0.40% to -0.36% in years 1-2 and -0.31% to -0.16% in years 4-5), and the calibration slopes were significantly less than 1 at all timeframes (calibration slopes: 0.05 to 0.66, $p < 0.012$ across years).

Thus, ClinValAI enabled the establishment of an effective cloud infrastructure to successfully perform the clinical validation of Mirai on a large and diverse dataset to study its generalizability.

2.5 Large-scale comparative evaluation of commercial FDA-approved AI Algorithms

The rigorous external validation of Mirai on the UW Medicine dataset demonstrated the feasibility of our framework in understanding the generalizability of deep learning models for medical imaging applications.

As we described at the start of this chapter, it is usually not feasible for a single institution to validate multiple commercial AI systems with its own data. Now that we have built ClinValAI and demonstrated its feasibility with Mirai, the next step is to show that ClinValAI can be used with a wide range of AI algorithms. Below, we describe the process of customizing ClinValAI to perform a large-scale comparative evaluation study of multiple FDA-approved commercial AI algorithms for mammography-based breast cancer detection using over 18,000 screening mammography exams contributed by seven U.S. regional registries affiliated with the Breast Cancer Surveillance Consortium (BCSC) [69]. BCSC registries systematically capture long-term cancer outcomes for each exam through linkage with local, regional, and state-level cancer registries, providing an unparalleled, gold-standard dataset.

We established a centralized cloud-based infrastructure on AWS (Amazon Web Services) using our ClinValAI framework to generate exam-level AI scores. Moreover, we implemented a secure mechanism for the BCSC registries to share de-identified data and customized our infrastructure to ensure robust external validation workflows.

2.5.1 Data Aggregation

We set up an S3 (Simple Storage Service) bucket and created individual folders for each BCSC registry. We leveraged the Secure File Transfer Protocol (SFTP) to implement an encrypted method that enabled registries to upload de-identified mammograms and associated manifest files from their databases. Every registry was provided with an authorization key that was configured to allow each registry only to access its corresponding folder in the cloud storage.

Additionally, for registries that did not have mammography exams stored in a separate database, we created an alternate mechanism that enabled them to transfer exams directly from their PACS (Picture Archiving and Communication System) server. We implemented a PACS-to-PACS transfer mechanism that allows interoperable sharing of DICOM images. Nearly all medical institutions today utilize PACS-to-PACS transfer for HIPAA-compliant communication. To support PACS-to-PACS transfer in our infrastructure, we deployed Orthanc[70], a lightweight, open-source PACS, on EC2. Orthanc communicates with an allowed list of outside senders to receive DICOM data, store it in the cloud, and stage it for later extraction, similar to SFTP.

2.5.2 Registry-specific customizations to the Data Ingestion Workflow

The data ingestion workflow consists of mechanisms for input auditing and data standardization. No additional customizations in the input auditing mechanism to check for corrupted files, missing DICOM headers, images from unsupported manufacturing devices, etc., were necessary.

However, we made multiple registry-specific modifications to the data standardization mechanism using ClinValAI's configurability. For example, different dataset batches from various registries required distinct imputation logic to infer laterality and projection information for images with missing DICOM headers. We added a registry-specific data restructuring feature to the data standardization mechanism since datasets received from all registries were organized in different formats. For example, data from some registries included additional image-level folders containing individual DICOMs from an exam. In contrast, other registries transferred DICOMs without any exam-level structuring, requir-

ing analysis of the DICOM headers to organize images into exam-level folders. We organized all datasets into a standardized exam-level format, thereby streamlining the implementation of subsequent algorithm-specific data restructuring pipelines. Such customizations helped expand the utility of ClinValAI.

2.5.3 Algorithm-specific customizations to the Scoring Workflow

The scoring workflow comprises mechanisms for data restructuring, algorithm inference, and score extraction. All AI algorithms had disparate input structuring requirements. Scoring workflows are algorithm-specific – essentially, all three mechanisms were customized for each commercial AI model. Similarly, each AI algorithm had its specific way of representing its outputs. For each AI model, data restructuring mechanisms were tailored to transform the datasets from the standardized format to the model-specific format. Additional criteria, such as the permitted number of images per view or the segregation of 2D and 3D images, were also considered while customizing the Docker images associated with the data restructuring mechanism.

AI algorithm inferencing mechanisms were configured to leverage and customize the Docker images comprising the deep learning model, programmatic dependencies, and the associated codebase necessary to process exams. For commercial algorithms with separate Docker images for scoring 2D and 3D images, the inference mechanism was split into two corresponding parallel pipelines. ClinValAI's feature of using the polling mechanism was also incorporated where applicable. Appropriate compute instances with multiple GPUs (Graphics Processing Units) based on the parametric complexity of the algorithms were specified in the framework and utilized by the infrastructure for executing scoring scripts.

The output extraction mechanism for each AI model was customized based on the model's generated files, for example, comma-separated values (CSV) documents, JavaScript Object Notation (JSON) objects, or DICOM Structured Reports (SRs). We tailored the pipeline to extract exam-level scores from output files based on the documentation shared by the AI vendors – essentially, score value from row-column index from CSV files, key path from JSON objects, and Concept Name Code Sequence Attribute from DICOM SRs. We also stored all algorithm-generated supplementary files, including logs and models' annota-

tions on images, for improved downstream interpretability and troubleshooting.

These algorithm-specific customizations to the scoring workflow streamlined our external validation efforts.

2.5.4 Troubleshooting-related customizations to the Output Processing Workflow

The output processing workflow comprises an output auditing mechanism and an output aggregation mechanism. ClinValAI's output auditing mechanism identifies exams with missing scores and helps determine issues with such exams. We performed additional troubleshooting steps to ensure that scores are received for the largest possible number of exams, thereby maintaining a significant sample size with a robust demographic distribution. For every AI algorithm, all exams with missing scores were passed again through the scoring workflow to check for any infrastructure issues. We parsed all workflow-specific logs to identify any errors associated with limited memory, insufficient compute, or problems with GPU drivers. After addressing these errors and re-executing the scoring workflow, the algorithm-generated logs were aggregated for any subsequently remaining exams with missing scores. We shared those logs with the AI vendors to understand potential issues with the DICOMs and scoring pipelines. By working in a closed feedback loop with the vendors, we updated the scoring scripts and resolved problems with DICOM metadata.

Incorporating these customizations into ClinValAI's output processing workflow enabled a systematic troubleshooting approach, providing us with exam-level AI scores for a large and diverse dataset.

2.5.5 Analysis

Patient Cohort

The BCSC study cohort consisted of 18,317 screening mammography exams from seven U.S. regional registries affiliated with the BCSC. All mammograms were from women over 40 years at the time of screening. We included bilateral screening mammography exams without a personal history of breast cancer, and with a two-year follow-up. All exams had either an initial end-of-day (EOD) BIRADS assessment—defined as a BIRADS assessment after a same-day work-up —or a final BIRADS assessment—defined as a BIRADS assessment after all work-up is completed, considering the full screening process, including any

diagnostic work-up, or both. The exams were sampled with a 1:2 case-control ratio: 3,512 exams had breast cancer detected within one year of screening, 2,262 had breast cancer detected between one and two years of screening, and the remaining 12,543 exams were controls (i.e., no breast cancer detected within the two-year follow-up period). Table 2.4 provides the distribution of 2D and 3D mammograms for the BCSC study.

Table 2.4: Distribution of 2D and 3D screening mammography exams.

Mammography Exam Type	All (n=18,317)	Control (n=12,543)	Case: <1 year (n=3,512)	Case: 1–2 years (n=2,262)
2D Digital Mammogram	14913 (81.4%)	10055 (80.2%)	3039 (86.5%)	1819 (80.4%)
3D Digital Breast Tomosynthesis	3404 (18.6%)	2488 (19.8%)	473 (13.5%)	443 (19.6%)

AI Score Distribution

To preserve the anonymity of the FDA-approved commercial AI algorithms participating in our large-scale comparative evaluation study, the deep learning models are denoted by variables A, B, and C. Figure 2.10 demonstrates the score distribution of these AI models across all exams, stratified based on the cancer status – control, case (1-2 years), case (<1 year). AI scores can be binarized by using the 90^{th} percentile of control exam scores as the threshold for classifying whether an algorithm detects cancer on a mammogram.

Qualitative analysis suggests that all three algorithms assign higher scores to exams with a cancer detected within one year of screening than to exams with a cancer detected within two years of screening. Similarly, for all three algorithms, scores assigned to controls are lower than those assigned to cases. Additionally, the distributions of scores are concentrated toward lower values for algorithm A and toward higher values for algorithm C.

Score distribution stratified by breast density

Analyzing the performance of AI algorithms based on breast density is essential to assess their generalizability. Figure 2.11 demonstrates the comparison of algorithms' performance for exams from dense (BIRADS c and d) vs. non-dense (BIRADS a and b) breasts. Qualitative assessment of the score distribution suggests that all AI algorithms demonstrate

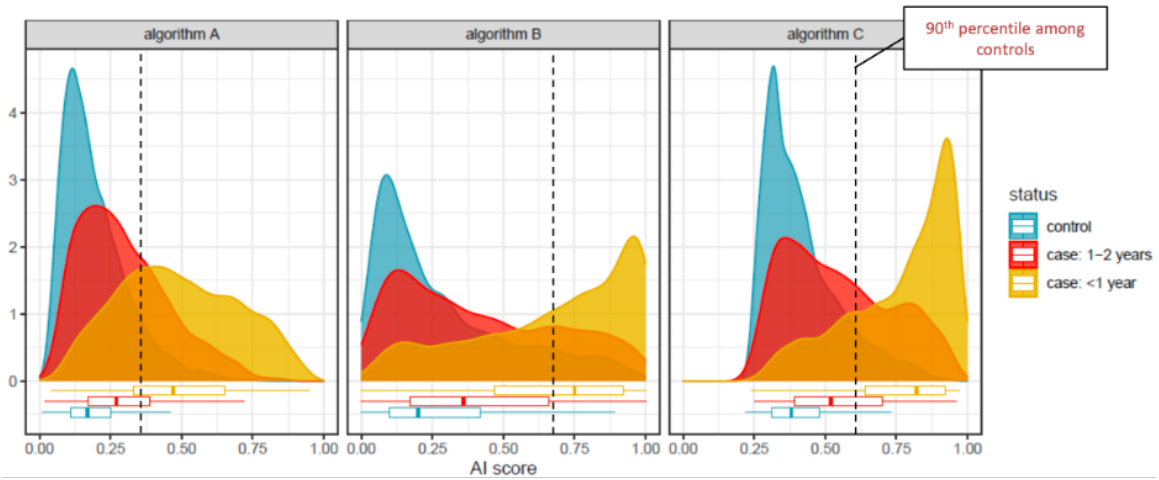


Figure 2.10: Distribution of exam-level scores from three FDA-approved commercial AI algorithms. The 90$^{\text{th}}$ percentile of scores among control exams is represented as a threshold to binarize AI scores.

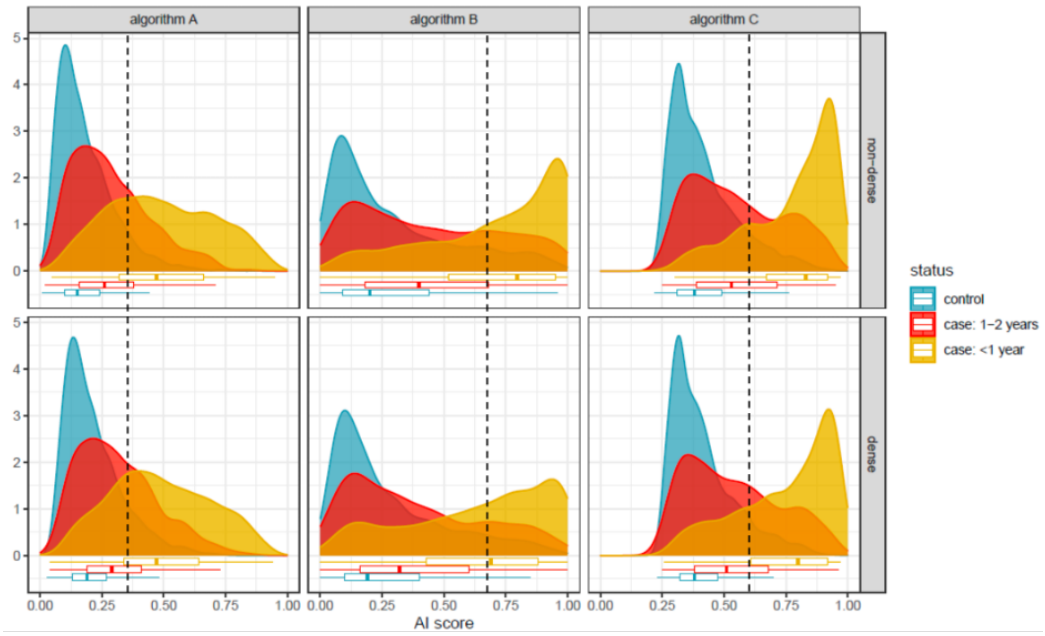


Figure 2.11: Stratification of AI algorithm scores based on breast density.

marginally improved classification of cases vs. controls for exams from non-dense breasts compared to dense breasts, highlighting the challenges of lesion detection in women with dense breasts. This observation is particularly evident for algorithms B and C, as the separation between the score distributions of cases and controls is more distinct in non-dense breasts than in dense breasts.

Score distribution stratified based on radiologist assessment

Analyzing the distribution of AI algorithm scores based on the initial end-of-day (EOD) and final radiologist assessments can help understand the potential for integrating AI algorithms into radiology workflows for breast cancer detection. To examine the benefits and concerns of AI-based breast cancer detection, Figure 2.12 and Figure 2.13 compare AI scores with binarized EOD and final BIRADS assessments, respectively. Observations are consistent in both scenarios. For exams with breast cancer detected within 1 year or between 1 and 2 years of screening, scores from all three AI algorithms show improved cancer detection, with a higher proportion of AI scores exceeding the detection threshold (defined as the 90th percentile of scores on the control exams).

Moreover, the proportion of exams deemed cancer-negative by radiologists but cancer-positive by AI algorithms demonstrates the potential of these models to detect cancer in exams where suspicious regions may not have been visible to radiologists. However, we also observe a higher proportion of false-negative AI assessments, especially for algorithms A and B, particularly for exams that develop breast cancer within one year of screening. For controls, while all AI algorithms show a higher proportion of scores below the detection threshold, the proportion above the threshold is higher than the radiologists' false-positive rates, underscoring the need for targeted fine-tuning efforts. We also observe several exams, in both cases and controls, where both radiologists and AI algorithms face challenges in interpreting those screening mammograms.

Thus, this analysis provides a method to compare the performance of AI algorithms with that of radiologists and to understand the potential advantages and caveats associated with the adoption of these FDA-approved models in radiology workflows.

Overall, this large-scale comparative evaluation study demonstrates the potential of

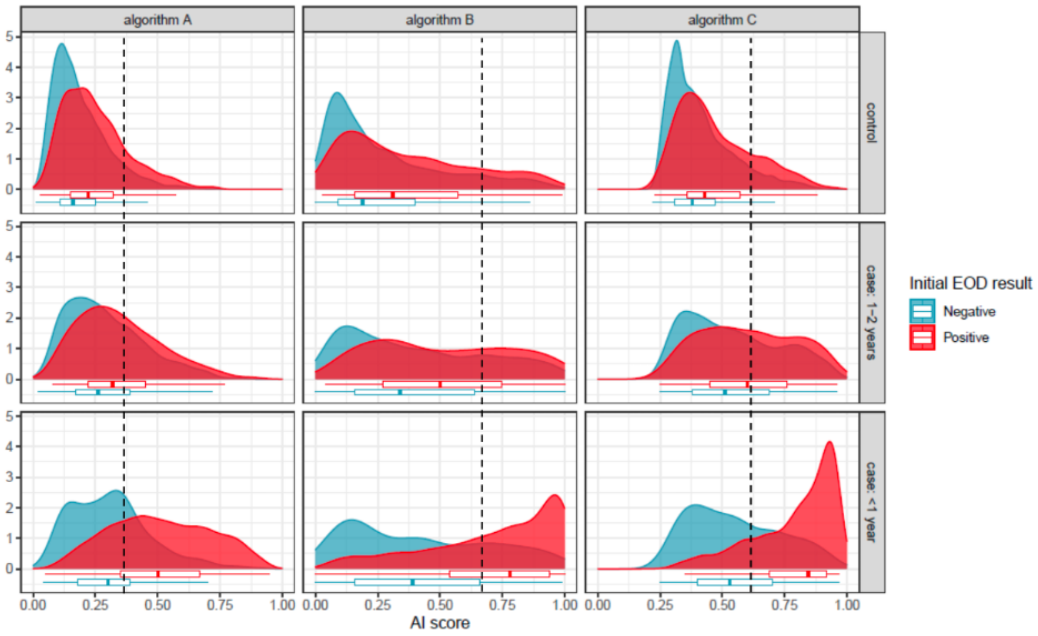


Figure 2.12: Comparison of AI score distribution and radiologists' end-of-day (EOD) assessments

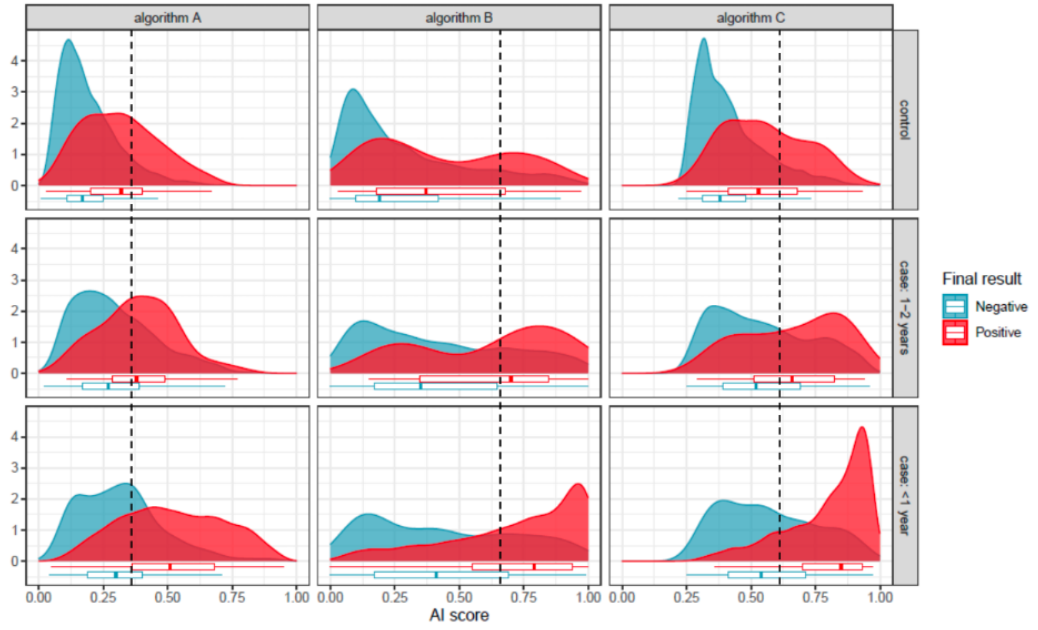


Figure 2.13: Comparison of AI score distribution and radiologists' final assessments.

leveraging ClinValAI to establish robust cloud-based infrastructures to assess the generalizability and real-world potential of AI algorithms before their integration into clinical workflows.

2.6 Clinical Impact

With multiple AI algorithms receiving FDA clearance for the same radiology applications, the selection and integration of these AI technologies into local clinical radiologists' workflow will become a major task for health care organizations. Successful and clinically meaningful deployment of AI will be contingent on the rigorous external validation of these models to select the most effective algorithm for a specific target population. Concerns regarding the underperformance of overparametrized deep learning models on out-of-distribution validation sets can lead to reservations about the generalizability of FDA-cleared AI algorithms. Thus, empowering healthcare organizations to conduct robust external validation of multiple commercial AI products locally before adoption in their target population will be crucial moving forward.

Administering external validation studies for imaging-based AI in healthcare algorithms should be a collaborative effort involving clinical leadership and interdisciplinary teams with multiple stakeholders possessing skills in IT infrastructure development, medical image analysis, AI, and biomedical informatics. Institution-specific validation infrastructures can be reused to periodically monitor the performance of AI algorithms over time, allowing the identification of AI model drift in target populations. Periodic revalidation of AI models eventually deployed in clinical workflows can identify temporal performance degradation for specific sub-populations. Thus, the continuous monitoring feature enables analyzing variations in model performance vis-à-vis data drift and model drift.

Detecting unexpected behavior in AI models through comprehensive error and bias analysis, and communicating validation results back to AI vendors, can also advance algorithm refinement. Essentially, model developers can leverage feedback from single-institution validations to establish targeted fine-tuning of deep learning algorithms and enhance the accuracy of AI models for specific subpopulations. Similarly, AI vendors can take necessary

steps to account for specific manufacturing devices if the captured images are not supported or cause reduced model performance. Independent external validation of commercial AI algorithms can thus motivate developers to enhance the explainability of their AI models, thereby enhancing the potential of receiving clinicians' trust. Overall, external validation studies can promote the development of robust, reliable, and trustworthy models, and foster enhanced academic-industry partnerships.

In addition to enabling individual healthcare institutions to perform external validation of promising AI algorithms (section 2.4), ClinValAI can be used to conduct large-scale comparative evaluations of multiple FDA-approved commercial AI tools (section 2.5) using medical imaging data acquired across diverse health systems. Such comparative studies can enable a better understanding of performance variation across AI models, enable stratified analyses across clinically meaningful subgroups, and evaluate the potential benefits and risks of integrating AI algorithms into clinical workflows. Moreover, by enabling large-scale external validation efforts on data from diverse cohorts, our work has the potential to foster health equity and overcome health disparities by promoting the development of robust, interpretable, and generalizable AI algorithms for healthcare applications.

Thus, our ClinValAI framework promotes the external clinical validation and comparative evaluation of AI algorithms on medical imaging exams, thereby providing the opportunity to reliably understand the real-world performance of AI-based clinical decision support systems in healthcare settings and their impact on patient care, health, and safety. ClinValAI can be leveraged to evaluate the generalizability of deep learning models on healthcare data from diverse demographics to analyze the differences in performance across various sub-populations and identify biases. ClinValAI can facilitate the detection of models' failure modes and enable an understanding of AI's potential to function as a standalone tool for diagnostic applications.

Overall, ClinValAI can pave the way for studying the capabilities of AI algorithms in optimizing clinical workflows and reducing the burden on the medical practitioners. By providing a foundational framework for establishing cost-efficient and robust infrastructures for external validation of AI algorithms, ClinValAI can help institutions adopt high-performing AI algorithms into their radiology workflows and promote improved outcomes.

Chapter 3

FEATURE-SPACE OCCLUSION TO EXPLAIN A DEEP LEARNING BREAST CANCER RISK PREDICTION MODEL

In Chapter 2, we demonstrated our work on addressing the challenges associated with performing external validation of deep learning algorithms for medical imaging applications. ClinValAI can help identify robust, generalizable AI tools. However, realizing the full potential of such black-box deep learning models requires understanding the factors driving their predictions. In this chapter, we present our work on providing clinically meaningful explainability to the risk estimates of the Mirai [26] algorithm for mammography-based breast cancer risk prediction. Various external validation studies [41–44], including our work in Chapter 2, have demonstrated Mirai's accurate and generalizable performance.

Here, we introduce an explainability technique that remains faithful to Mirai's architecture, with the aim of understanding the mammogram imaging features driving its risk predictions, guiding timely interventions, and assessing the trustworthiness of model outputs. The following sections are from our manuscript titled 'Feature-space Occlusion to Explain a Deep Learning Breast Cancer Risk Prediction Model', currently under review at Radiology: Artificial Intelligence journal, with authors Ojas A. Ramwala, Cody Schopf, Kathryn P. Lowry, John H. Gennari, Sean D. Mooney, Christoph I. Lee, and William Lotter.

3.1 Introduction

Advancements in artificial intelligence (AI) for medical imaging have led to the development of promising mammography-based algorithms for breast cancer risk assessment [25, 40, 71, 72]. These deep learning models, which predict the risk of future breast cancer based solely on image features, have been shown to outperform traditional clinical risk assessment tools [25, 47, 73, 74]. In particular, Mirai [26], an open-source mammography image-based deep learning model for predicting five-year breast cancer risk, has demon-

strated accurate and generalizable performance across diverse subpopulations in large-scale validation studies [41–44]. These advancements have garnered much enthusiasm as a path toward personalized screening strategies. For example, high-risk women may benefit from more intensive screening, including supplemental imaging, whereas low-risk women could potentially undergo longer screening intervals.

However, the specific underlying mechanisms of mammography-based AI breast cancer risk prediction remain unclear. One possibility is that models like Mirai are identifying localized imaging features, such as the early signs of a developing lesion. Conversely, the model may make predictions based on global features like breast density or texture patterns. Clarifying the features driving predictions is essential for optimizing clinical utility and advancing the understanding of breast cancer risk more broadly. If the risk features can be localized, focused diagnostic imaging and short-interval imaging follow-up may be especially beneficial. Explaining Mirai's predictions, and image-based models more generally, presents numerous challenges. Traditional explainability methods like Grad-CAM [75] have been shown to have limited utility in representing the decision-making process of AI algorithms, especially regarding localization [76, 77]. Studies specifically aiming to explain Mirai have instead focused on how certain subsets of features associate with its predictions or performance. AsymMirai simplifies components of Mirai's architecture to explicitly compute bilateral dissimilarity in the feature space [78]. Recently, Wang et al. [79] identified that some of the features extracted by Mirai appear to quantify the presence of calcifications and masses, and the values of these features correlate with Mirai's predictions.

In this study, we developed an explainability technique that quantifies the influence of localized regions of interest (ROIs) on Mirai's risk estimates without altering its architecture. The approach masks ROIs in Mirai's feature space, leveraging the model's use of Global Max Pooling to explain its predictions more faithfully. By annotating a screening cohort with both index (cancer diagnosed) and negative prior exams, we assessed the impact of removing current and future cancer regions on Mirai's risk scores and classified its predictions into different semantic categories. Our technique and analysis aim to inform both technical and clinical stakeholders, including radiologists, by interpreting Mirai's predictions and understanding whether local or global imaging features drive its risk estimates.

3.2 Materials and Methods

3.2.1 Patient Cohort

The study cohort originated from a consecutive set of 2D screening digital mammograms acquired between 2010-2014 from four imaging facilities within an academic health system. Mirai's performance has previously been validated in this patient population [43], demonstrating generalization consistent with its performance in other cohorts. Five-year cancer outcomes were obtained for each woman through linkage to the state cancer registry. From the original consecutive cohort, we included exams in this study from women with a diagnosis of breast cancer (invasive or DCIS) within five years after the screening exam date. Furthermore, we required that each woman have both an index screening mammogram, defined as the mammogram for which breast cancer is detected in clinical practice within 90 days of screening, and a prior negative screening mammogram within the preceding five years. Given the extensive annotation process described below, we subsampled the eligible exams to a target of > 200 and prioritized exams from women with multiple priors to enable longitudinal analysis. All data were de-identified prior to analysis in compliance with HIPAA. The study was conducted under a waiver of consent with approval granted by the institutional review board.

3.2.2 Annotation Strategy

A fellowship-trained breast radiologist annotated each of the included exams (Figure 3.1). Rectangular bounding boxes were drawn on index exams around the malignant lesion(s) on the CC (Cranio-Caudal) and MLO (Medio-Lateral Oblique) views of the cancer-affected breast. Then, regions of matching size and location were annotated on the CC and MLO views of prior exams to denote the regions of future lesion development. Additionally, the radiologist annotated anatomically matched regions in both views of the contralateral breast for each exam (i.e., the breast where cancer did not develop). We refer to the annotations in the cancer-affected breast as "case" annotations and those in the contralateral breast as "control" annotations.

For each case annotation, the radiologist also indicated the visibility of the lesion (yes or no), lesion type (mass, calcification, architectural distortion, and/or asymmetry), and other

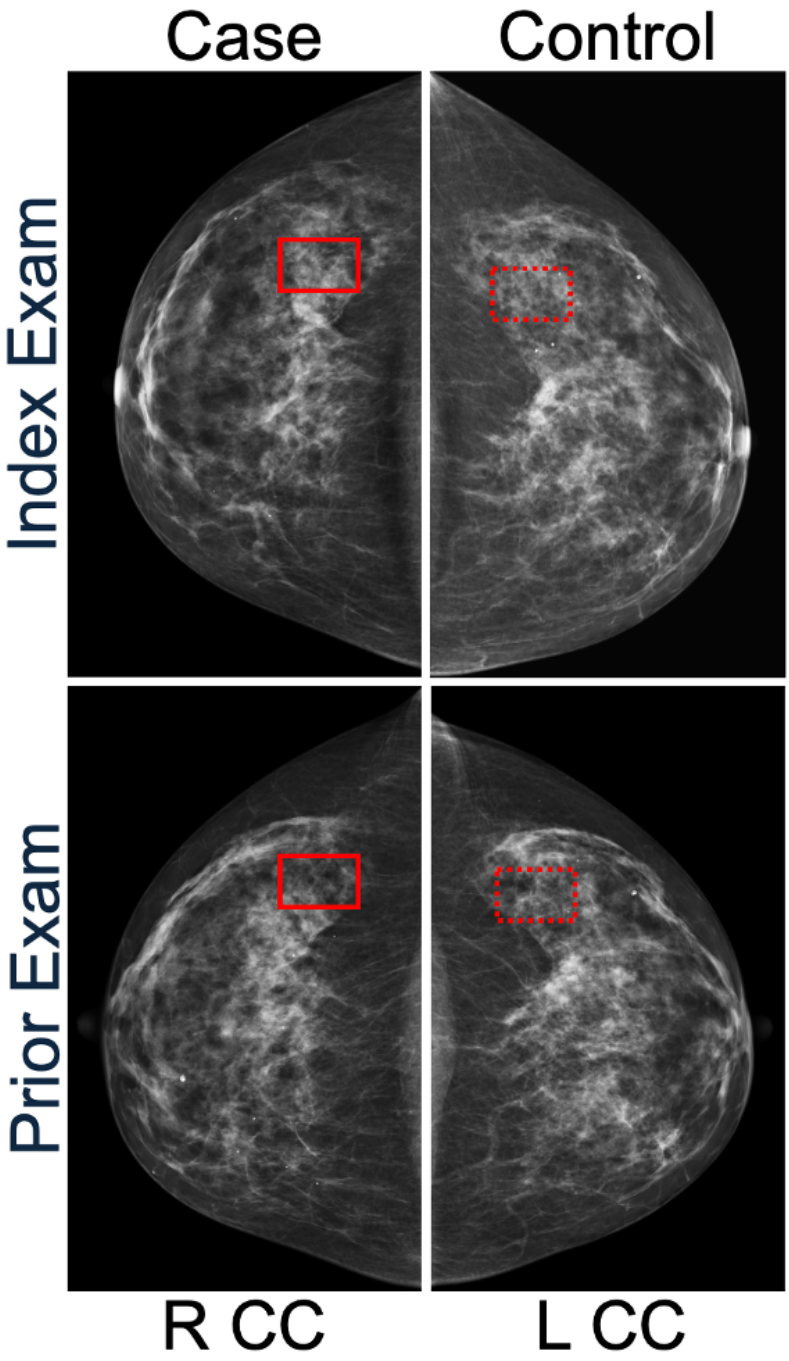


Figure 3.1: Annotation Strategy. A bounding box denoting a malignant lesion is drawn on the index exam, and a region of matching size and location is drawn on the prior exam. These are referred to as "Case Annotations". "Control Annotations" are provided by drawing bounding boxes on anatomically matched regions in the contralateral breast of the index and prior exams.

associated findings (nipple retraction/skin and trabecular thickening/lymphadenopathy). Annotations were made using a customized version of the VGG Image Annotator tool [80]. For all analyses, all bounding box annotations were resized to a standardized 512x512 pixels, covering the full range of original bounding boxes and approximating breast quadrant-level localization based on Mirai's input image dimensions. This more standardized localization scale was chosen to reflect potential future clinical utility; localizing a developing lesion to a breast quadrant would facilitate targeted diagnostic imaging, whereas less specific localization would not. The lesion visibility and type annotations were aggregated at the exam level for subgroup analysis. An exam was classified as having a visible lesion if any view contained a visible lesion annotation. Exam-level lesion type was defined as the union of lesion types annotated across all views. Lesion-type analyses were omitted for exams without visible lesions.

3.2.3 Overview of Mirai's Deep Learning Architecture

Details of Mirai's deep learning architecture have been previously described [26]. Briefly, Mirai processes the four standard images of a 2D digital mammogram – CC and MLO views of the left and right breasts. Each image is first processed individually by the Image Encoder module, consisting of a ResNet-18 [81] network that outputs a feature matrix of size $[512, 64, 52]$, representing the channels, height, and width, respectively. The feature matrix is passed to a Global Max Pooling layer, a common operation in deep learning models like ResNet [81], that extracts the maximum value for each of the 512 feature channels across space, generating a $[512, 1]$ dimensional vector. The vectors for each of the four views are then passed to the Image Aggregator module, a Transformer [82] network, to generate a combined representation, followed by a Risk Prediction module to generate 5-year absolute risk predictions.

3.2.4 FOCUS: Feature-space OCclusion for Understanding Saliency

To assess whether Mirai focuses on regions where cancer develops, an intuitive approach would be to 'remove' these regions from the input image (i.e., by setting the pixels to zero) and observe the change in risk scores. However, such perturbations would render the images out of distribution and may not faithfully represent changes in risk predictions.

Instead, we developed the FOCUS (Feature-space OCclusion for Understanding Saliency) method (Figure 3.2), which masks specific regions in the feature space rather than the pixel space. Our approach leverages Mirai's use of Global Max Pooling. First, coordinates in pixel space are mapped to feature space by accounting for all neural network operations of the ResNet-18 architecture of Mirai's Image Encoder. Then, to mask a specific region on a mammogram image, the features with coordinates that overlap with that region are occluded (set to negative infinity). The modified features are then passed to Mirai's Global Max Pooling layer and Mirai's subsequent modules. Since the Global Max Pooling layer collapses the feature matrix into a vector by selecting the maximum value for each feature channel across space, the occlusion of features from a specific region forces the model to 'focus' exclusively on other regions while maintaining the fidelity of features from these regions (Figure 3.3).

3.2.5 Influence of Cancer Regions on Mirai's Risk Assessments

For each mammogram, we computed Mirai's risk predictions when applying FOCUS to case annotations ('case-occlusion') and, separately, to control annotations ('control-occlusion'). In both scenarios, the occlusion was performed simultaneously in the CC and MLO views for the corresponding breast (Figure 3.4). The effect of occlusion was quantified by comparing the resulting 5-year risk score to Mirai's original 5-year risk score. Furthermore, using Mirai's reported high-risk threshold of 2.6% [26], we assessed whether the case or control occlusion changed the risk estimates of high-risk exams to low-risk (defined as a risk prediction below the high-risk threshold).

3.2.6 FOCUS Maps: Visualizing Mirai's Feature Importance Distribution

For predicted high-risk exams that did not shift to low-risk upon case-occlusion, we performed additional analysis to understand the features driving the high-risk prediction. To do so, we developed visualizations using our FOCUS technique to quantify the importance of each image region on Mirai's predictions. Based on Mirai's input size of 2048x1664 pixels, we sampled 256x256 pixel patches with a stride of 128x128 pixels, resulting in 180 patches per view, or 720 patches per mammogram. We applied FOCUS separately to each patch and calculated the difference between the occlusion-based and original risk scores, termed

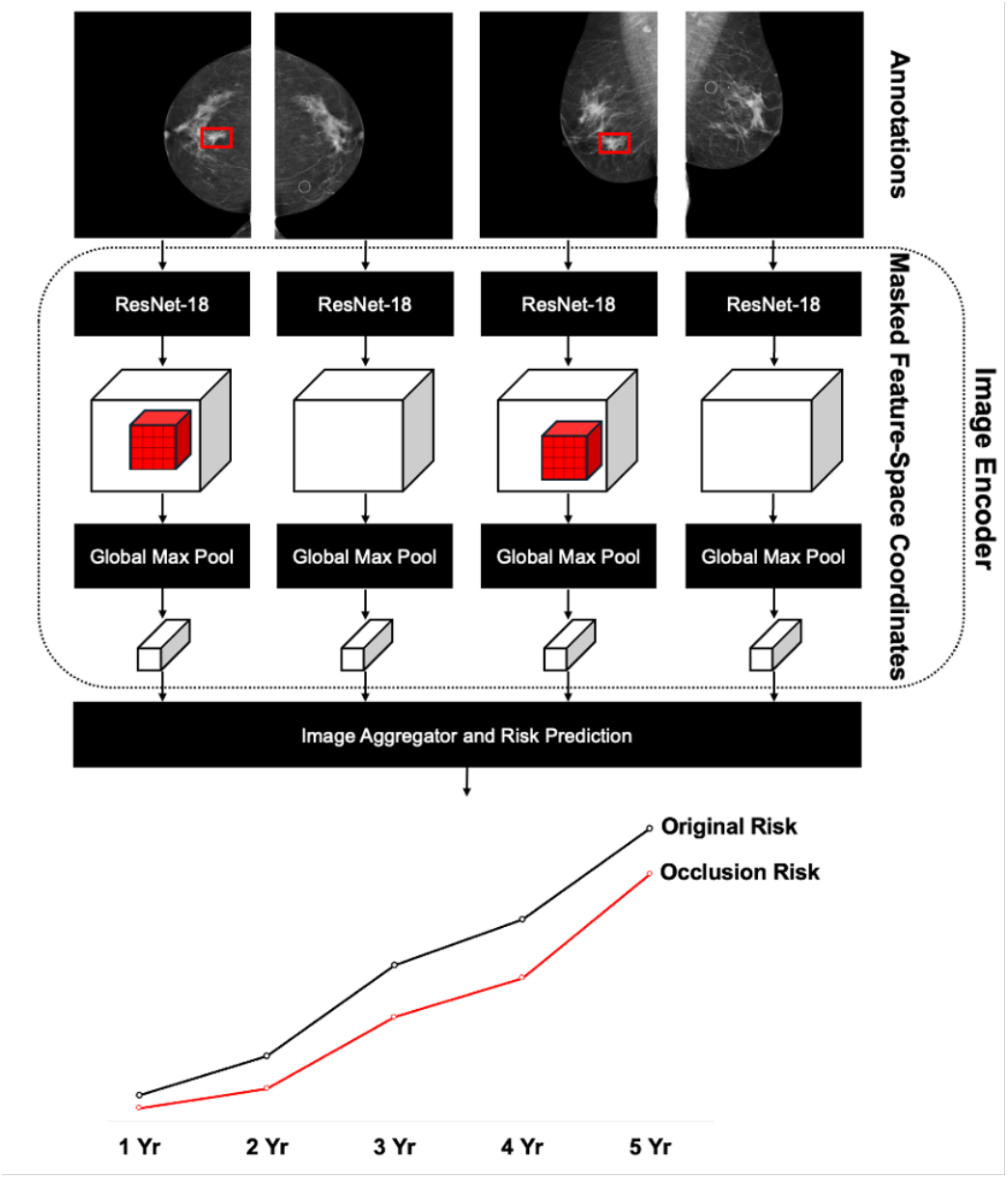


Figure 3.2: FOCUS method applied to Mirai. Each pixel of a mammogram is mapped to coordinates in the feature space based on the architecture of Mirai's image encoder. To remove annotated regions from Mirai's interpretation, the corresponding coordinates in ResNet-18's feature matrix are masked, i.e., the voxels in feature space indicated by the red cubes are set to negative infinity. This ensures that the Global Max Pooling layer selects features from other regions when generating the feature vector for that view. The risk score generated is then compared to the original risk score to estimate the impact of the annotated region on Mirai's prediction.

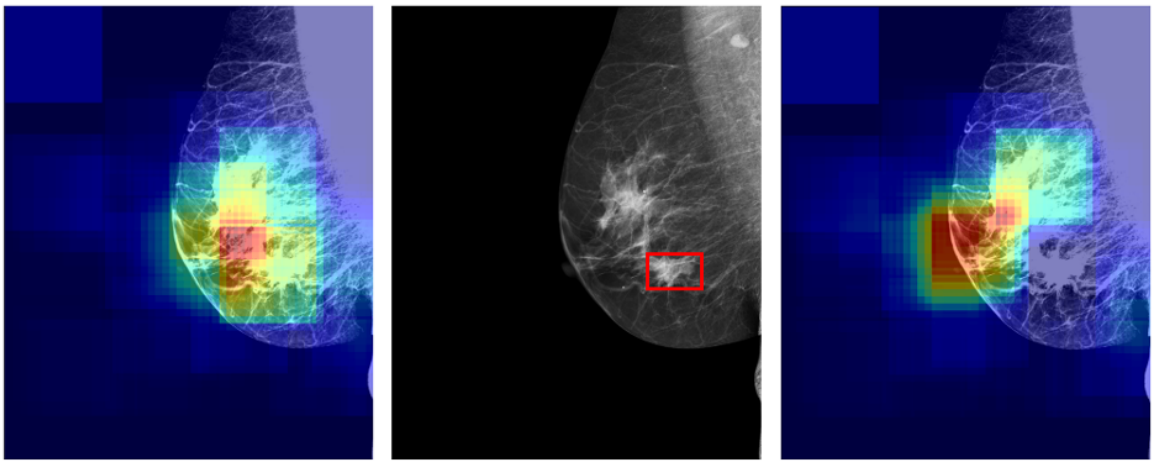


Figure 3.3: Validating the implementation of the FOCUS technique. (a) Original heatmap developed based on the projection of the coordinates of the maximum value from each channel of the output feature vector of Mirai's Global Max Pooling layer to the input image. The intensity of each region denotes the proportion of feature space channels that receive maximum information from that patch. (b) Input mammogram view with a red bounding box indicating the Radiologist's annotation of the lesion patch. (c) Heatmap generated after performing feature-space occlusion to mask the lesion patch from Mirai's inference mechanism. The minimum intensity at the annotated region indicates that Mirai's interpretation from that patch was not passed to the subsequent layers and, therefore, was not considered in its risk prediction.

the FOCUS Score, reflecting the influence of the patch on Mirai's predictions. We visualized these patch-level FOCUS Scores as heatmaps, termed FOCUS Maps, and identified the patch with the maximum FOCUS Score per view and exam (Figure 3.5). As patches are sampled with a stride equal to 50% of the patch size, overlapping regions can occur across patches. For each overlapping region, we assigned the maximum FOCUS score among the contributing patches. To ensure meaningful analysis and visualization, only patches whose occlusion resulted in a reduction in risk scores and contain less than 50% background region (defined as having a pixel value of 0) are included when creating the FOCUS Maps.

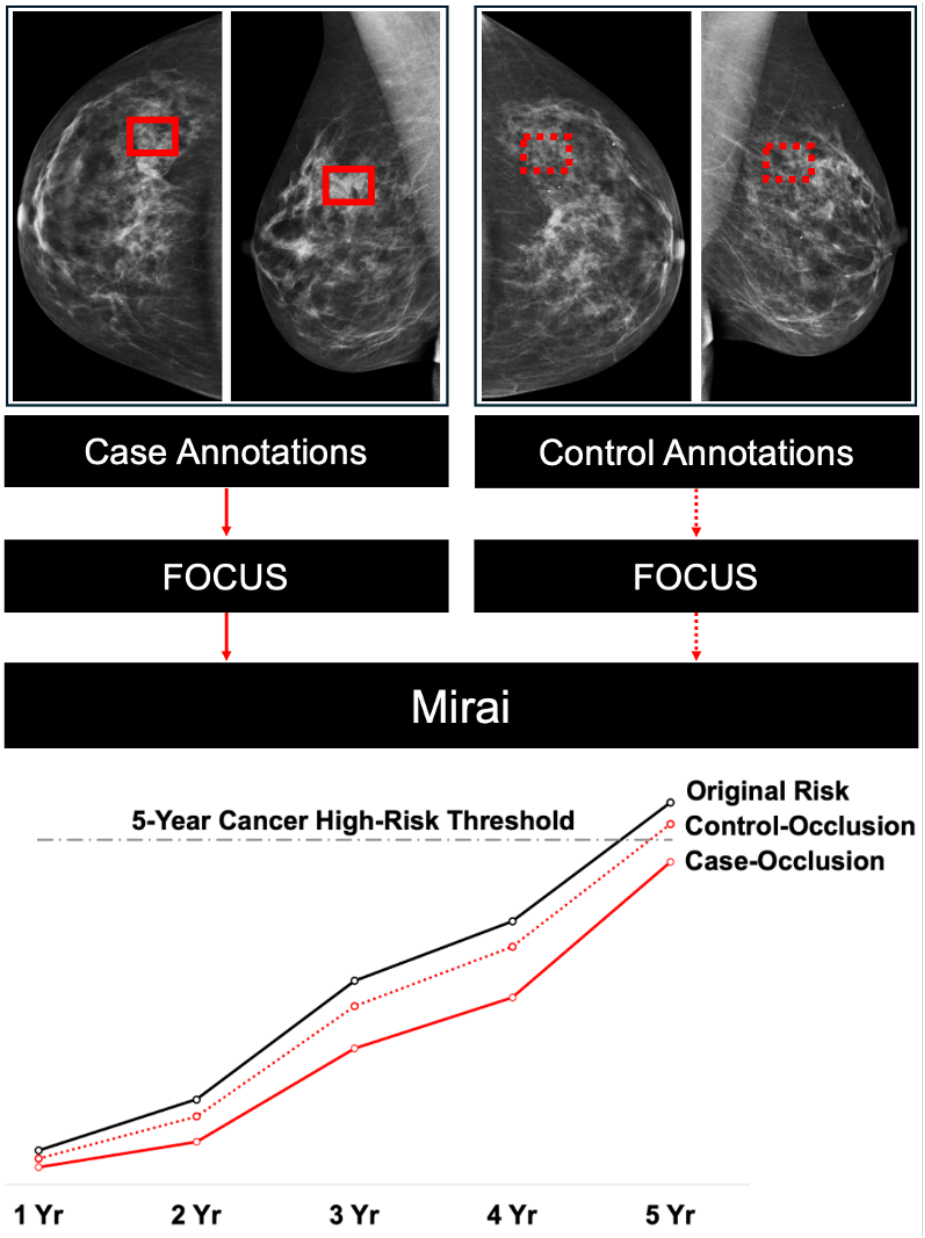


Figure 3.4: Illustration of case vs. control occlusion to assess the impact of cancer regions on Mirai's predictions. Case occlusion risk scores are computed by masking out the regions in the CC and MLO views of the breast where cancer was subsequently detected (case-annotations). Control occlusion risk scores are separately computed by masking anatomically-matched regions in the CC and MLO views of the contralateral breast (control-annotations). Occlusion-based risk scores are compared to the original risk scores to assess the influence of lesions on Mirai's predictions. Mirai's 5-year high-risk threshold is used to evaluate whether occlusions change Mirai's predictions from high-risk to low-risk.

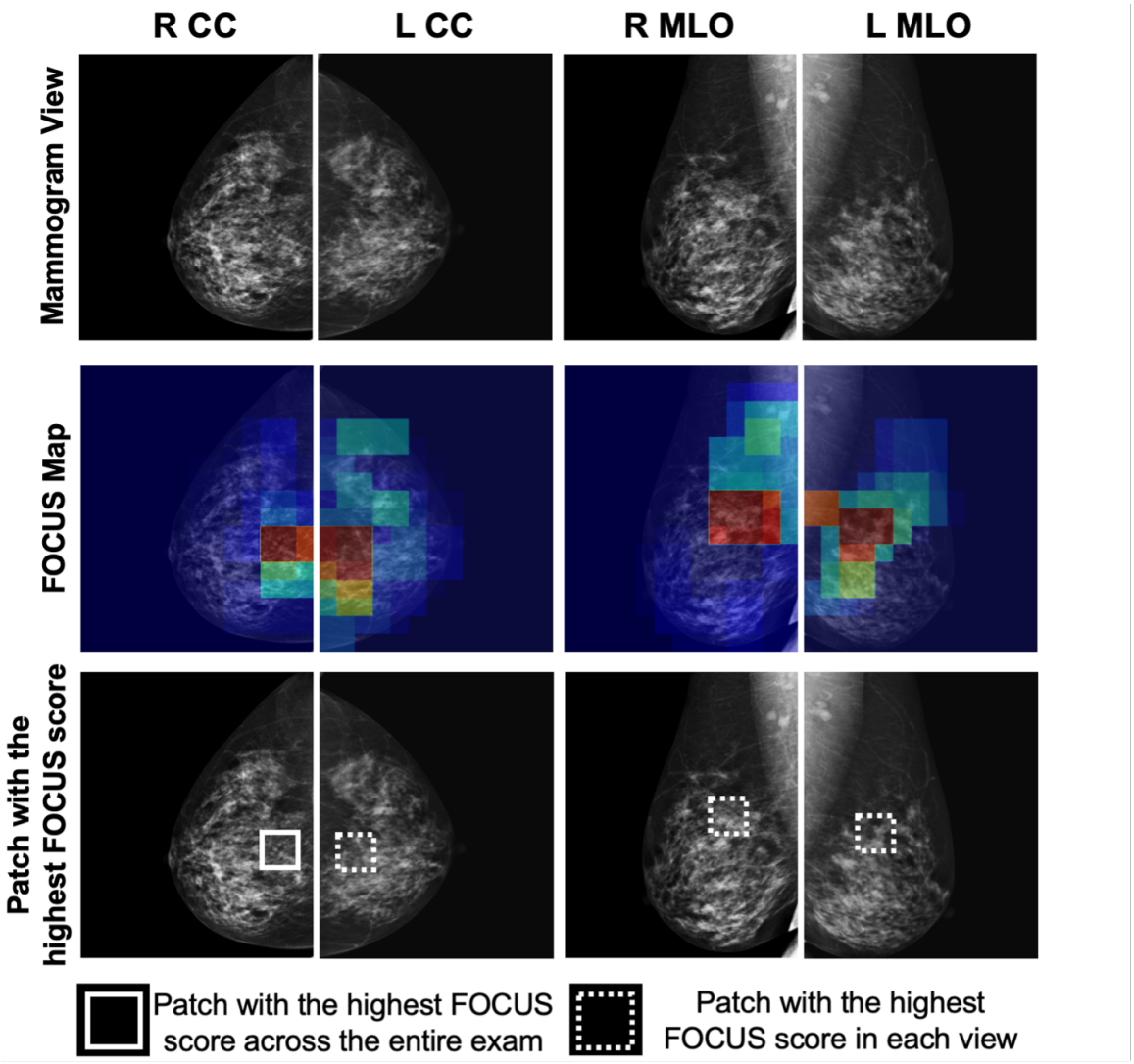


Figure 3.5: FOCUS Map to visualize Mirai's feature importance distribution. Image patches are iteratively masked to generate a FOCUS Score for each patch in each view. The patches with the highest FOCUS Score in each view and across the entire mammogram are identified for further analysis.

3.2.7 Categorization Scheme

Exams remaining high-risk upon case-occlusion were classified into distinct categories based on the FOCUS scores (Figure 3.6). The first level of categorization pertained to whether the risk assessment was driven by global or localized features. If the highest FOCUS Score

for the exam was ≥ 0.01, i.e., occluding a single patch resulted in a 1% decrease in the 5-year predicted risk, the prediction was considered localized. This threshold was selected based on the approximate difference between the high-risk threshold of 2.6% and the median predicted risk across the study population of 1.5% (see Results). In other words, if masking a 256x256-pixel region ($\sim 0.5\%$ of the mammogram) reduced the risk score by an amount comparable to the difference between a high-risk and typical mammogram, then localized features were deemed to meaningfully influence Mirai's prediction. If none of the patches in a mammogram had a FOCUS Score that crossed this localization threshold, the risk assessment was said to be based on global features. Localized predictions were further subclassified according to whether the patch with the highest FOCUS Score belonged to the cancer-affected or contralateral breast, with exams labeled as "lesion hit" if the centroid of the patch fell within the case annotation and a "lesion miss" otherwise.

3.2.8 Qualitative Analysis

Exams that were not classified as "lesion hit" were reviewed by three fellowship-trained breast radiologists. Using the FOCUS Maps, the radiologists assessed whether there were common imaging patterns that seemed to be driving Mirai's predictions. A similar qualitative assessment was performed for prior exams in which case-occlusion altered Mirai's prediction from high-risk to low-risk.

3.2.9 Statistical Analysis

Occlusion-based changes in risk scores were quantified using the mean and interquartile range (IQR). The difference in risk reduction between paired case and control occlusions was tested for statistical significance using the Wilcoxon signed-rank test [83], as it is non-parametric and does not assume a normal distribution of data. We considered a two-sided P value < 0.05 to determine statistical significance. Statistical analyses were performed using Python (version 3.7.3, https://www.python.org/) and the SciPy (version 1.7.3, https://scipy.org/) package. We implemented the FOCUS technique within the ClinValAI [43, 45] framework to establish an efficient, cloud-based model explainability infrastructure for Mirai and performed all experiments by customizing its workflows.

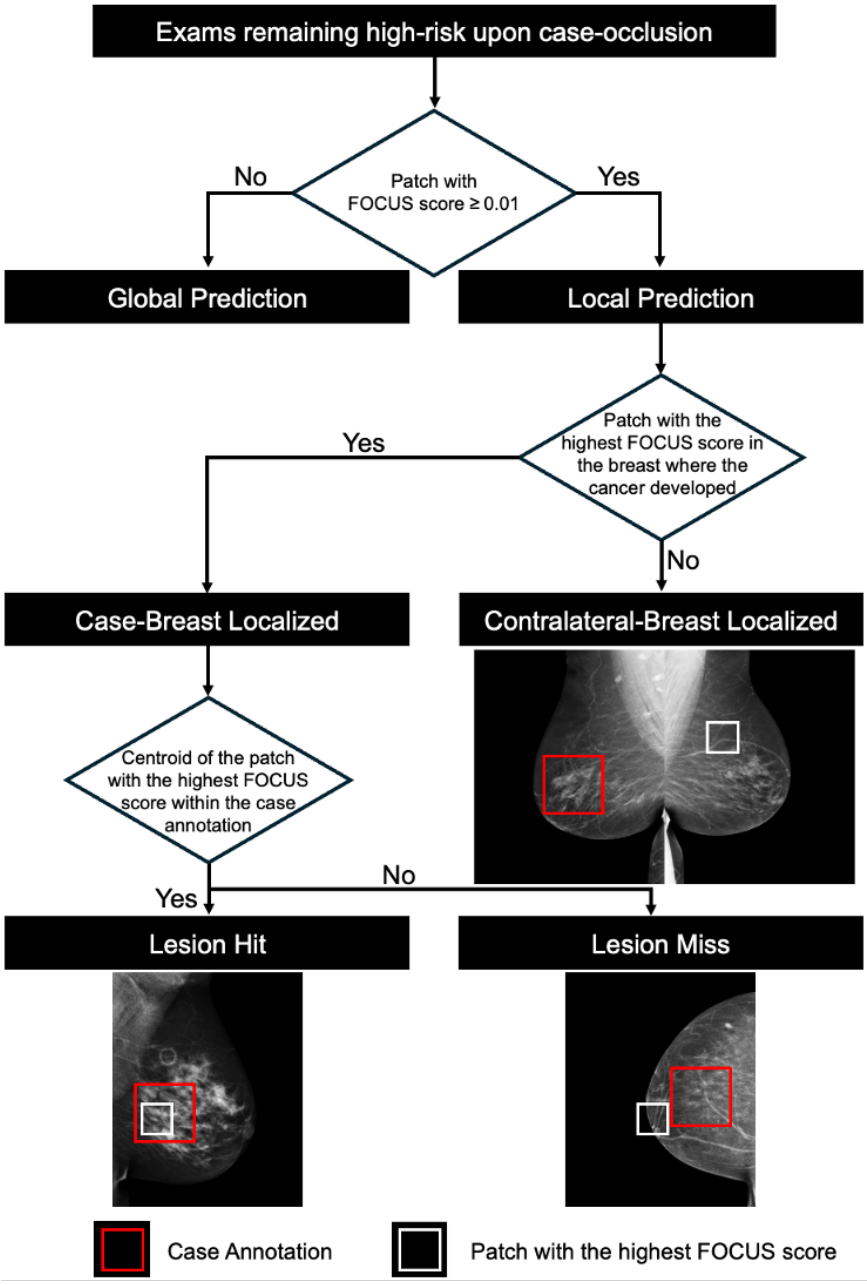


Figure 3.6: Categorizing exams that remain high-risk upon case-occlusion. Exams predicted as high-risk by Mirai, even after case-occlusion, were characterized. Risk assessments were classified as local predictions if any patch in a mammogram view had a FOCUS score $\geq$ 0.01 and as global predictions otherwise. Local predictions were further categorized based on whether the patch with the highest FOCUS score fell within the case or contralateral breast, and whether the centroid of the patch fell within the case annotation.

3.3 Results

3.3.1 Patient Characteristics

The retrospective consecutive cohort consisted of 72,983 screening mammograms from 33,373 women. Per Mirai's criteria, excluding exams of women with a personal history of breast cancer, exams of women of age < 40 or ≥ 80 years, exams of women with breast implants, and exams with fewer than the four standard screening views resulted in 57,690 exams from 27,911 women (Figure 3.7). Mirai's median and average 5-year risk predictions across the 57,690 exams were 1.51% and 1.98%, respectively ($IQR: 1.21\%, 2.07\%$). Retaining only mammograms from women with detected breast cancer with both index and prior exams available resulted in 354 exams from 127 women. Sampling to the target of 200 exams resulted in 220 exams from 70 women. After excluding exams from women with multiple lesion sites in the index exam (8 [3.8%] exams), which may confound the localization analysis, the final set included in analysis consisted of 212 exams from 68 women, with the index exam and all available prior screening exams within five years included for each woman in the subset (Table 3.1). All the women had unilateral breast cancer, even though bilateral breast cancer was not an exclusion criterion. Across the 212 exams, Mirai's average 5-year risk score was 4.83% ($IQR: 1.79\%, 4.98\%$), with an average of 8.01% ($IQR: 2.36\%, 13.47\%$) for the index exams and 3.33% ($IQR: 1.61\%, 2.90\%$) for the prior exams.

3.3.2 Effect of cancer regions on Mirai's risk predictions

Applying feature occlusion to the 212 exams resulted in a significantly higher score reduction for case-occlusion compared to control-occlusion ($p < 0.001$), with an average 5-year risk score change of -1.22% ($IQR: -0.78\%, 0.03\%$) for case-occlusion, compared to 0.06% (IQR: -0.16%, 0.20%) for control-occlusion (Table 3.2). This effect was significant for both index exams ($-2.68\%; IQR: -3.51\%, -0.01\%; p < 0.001$) and prior exams ($-0.53\%; IQR: -0.37\%, 0.03\%; p < 0.001$). Within prior exams subgroups, the effect was significant in both priors with visible lesions ($-0.86\%; IQR: -0.87\%, -0.02\%; p < 0.001$) and priors without visible lesions ($-0.45\%; IQR: -0.30\%, 0.04\%; p = 0.013$), and in priors within 2 years of the index exam ($-0.96\%; IQR: -0.51\%, 0.00\%; p = 0.001$) and priors before 2 years ($-0.26\%; IQR: -0.28\%, 0.07\%; p = 0.026$) (Table 3.3). The magnitudes of the effects

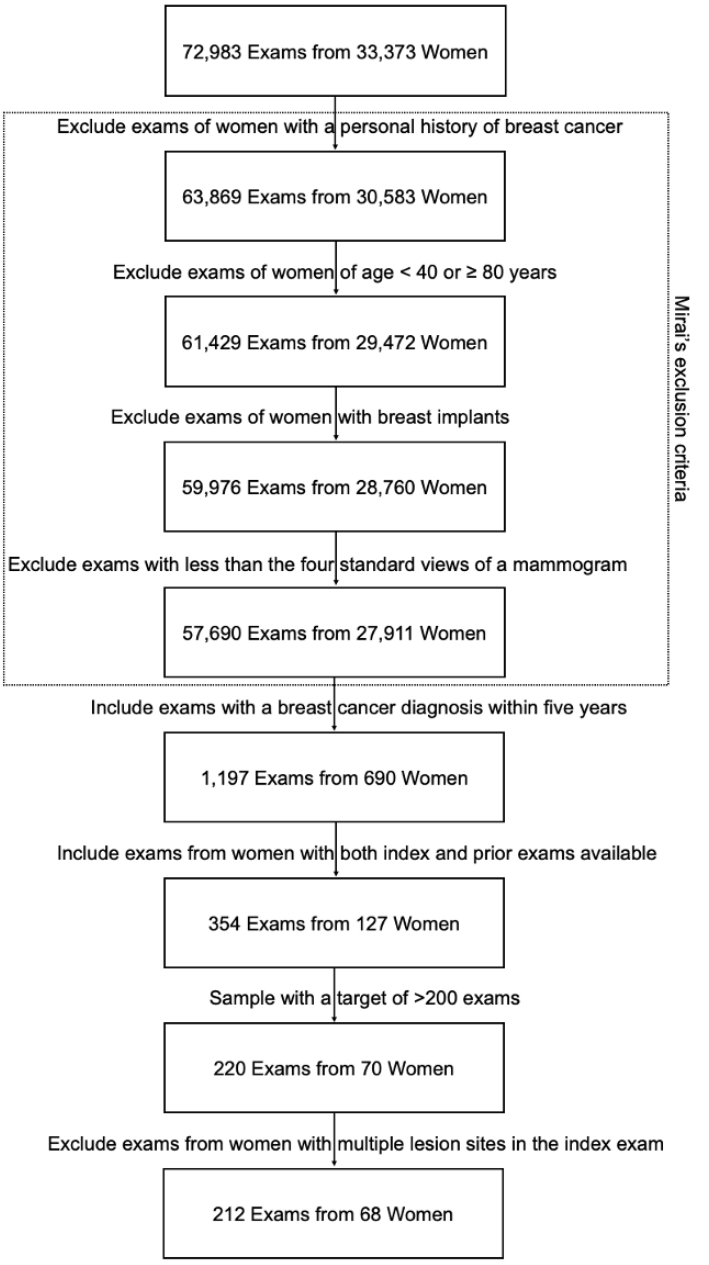


Figure 3.7: Study Cohort Selection. We applied Mirai's exclusion criteria based on the personal history of breast cancer, age, presence of breast implants, and the standard views of a screening mammogram. We included exams from women with a breast cancer diagnosis within five years with an available index exam and prior screening mammogram(s) within the preceding five years. After sampling a target of > 200 exams and excluding exams with multiple lesion sites in the index exam, our analysis cohort comprised 212 exams from 68 women.

Table 3.1: Patient and lesion characteristics across mammography exams.

Variable	All (n=212)	Index (n=68)	Priors (n=144)
Age			
40-49	27 (12.7%)	5 (7.4%)	22 (15.3%)
50-59	75 (35.4%)	21 (30.9%)	54 (37.5%)
60-69	76 (35.8%)	26 (38.2%)	50 (34.7%)
70-79	34 (16.0%)	16 (23.5%)	18 (12.5%)
Race/Ethnicity			
White	170 (80.2%)	54 (79.4%)	116 (80.6%)
Black	19 (9.0%)	5 (7.4%)	14 (9.7%)
Asian	12 (5.7%)	5 (7.4%)	7 (4.9%)
Hispanic	11 (5.2%)	4 (5.9%)	7 (4.9%)
Breast Density			
Not Dense (a and b)	94 (44.3%)	29 (42.6%)	65 (45.1%)
Dense (c and d)	117 (55.2%)	38 (55.9%)	79 (54.9%)
Unknown	1 (0.5%)	1 (1.5%)	0 (0.0%)
Initial BI-RADS Assessment			
0	92 (43.4%)	66 (97.1%)	26 (18.1%)
1	76 (35.8%)	0 (0.0%)	76 (52.8%)
2	43 (20.3%)	1 (1.5%)	42 (29.2%)
Unknown	1 (0.5%)	1 (1.5%)	0 (0.0%)
Lesion Visible			
Yes	97 (45.8%)	67 (98.5%)	30 (20.8%)
No	115 (54.2%)	1 (1.5%)	114 (79.2%)
Lesion Type*			
Calcification	42 (19.8%)	28 (41.2%)	14 (9.7%)
Asymmetry	34 (16.0%)	24 (35.3%)	10 (6.9%)
Mass	23 (10.8%)	16 (23.5%)	7 (4.9%)
Architectural Distortion	3 (1.4%)	3 (4.4%)	0 (0.0%)
N/A (Not Visible)	115 (54.2%)	1 (1.5%)	114 (79.2%)

*Two lesion types were observed in five exams.

were larger in index exams, priors with visible lesions, and priors within 2 years of the index exam. In index and prior exams with visible lesions, case-occlusion resulted in a higher risk reduction than control-occlusion for each lesion type (**??**), suggesting that Mirai does not simply cue on one type of lesion. Exams with masses showed the highest average original risk prediction (9.27%) and the largest average reduction with case-occlusion (−4.94%). Based on Mirai's high-risk threshold of 2.6%, 93 exams (48 index and 45 priors) were predicted as high-risk. Case-occlusion resulted in a significantly larger score reduction ($-2.63\%; IQR: -3.74\%, 0.00\%$) relative to control occlusion ($0.10\%; IQR: -0.13\%, 0.66\%$) in these exams ($p < 0.001$), including for both index exams ($-3.68\%; IQR: -4.78\%, 0.19\%; p < 0.001$) and prior exams ($-1.51\%; IQR: -2.24\%, -0.04\%; p = 0.003$). For high-risk prior exams, the effect was significant for those with visible lesions ($-2.11\%; IQR: -4.16\%, -0.36\%; p = 0.037$) and those without visible lesions ($-1.33\%; IQR: -1.28\%, -0.02\%; p = 0.020$). Case-occlusion changed Mirai's interpretation to low-risk for 29.0%(27/93) of the high-risk exams (33.3%[16/48] for index exams and 24.4%[11/45] for prior exams), compared to only 3.2% (0%[0/48] index and 6.7%[3/45] priors) for control-occlusion. For subsets of priors, the percentage change from high-risk to low-risk was $10.0\%[1/10], 28.6\%[10/35], 19.0\%[4/21]$, and 29.2%[7/24] for exams with visible lesions, no visible lesions, priors within 2 years, and priors before 2 years, respectively.

3.3.3 Categorization of exams remaining high-risk after case-occlusion

We applied our proposed categorization scheme (Figure 3.6) to the 66 high-risk exams (32 index, 34 priors) that did not change to low-risk upon case-occlusion. Seven (10.6%) of these exams (2 index, 5 priors) were classified as "global predictions" as they did not have a single patch with a FOCUS Score greater than the localization threshold, implying that the high-risk assessment was not strongly influenced by a single location for these exams (Table 3.4). Further analysis revealed that all seven global predictions occurred in women with dense breasts, compared to 55.9% (33/59) of exams categorized as "localized predictions".

Of the 59 localized predictions (30 index, 29 priors), 45 (76.3%) were localized to the cancer-affected breast compared to 14 (23.7%) in the contralateral breast. This trend was present in both index and prior exams, with index exams showing a larger effect (86.7%

Table 3.2: Effect of case and control occlusion on Mirai's risk predictions.

Mammography Exam Type	Exam Counts	Original Risk (mean, IQR)	Risk Reduction Case	Risk Reduction Control	p-value	Change to Low Risk* Case	Change to Low Risk* Control
All Exams	212	4.83% [1.79%, 4.98%]	−1.22% [−0.78%, 0.03%]	0.06% [−0.16%, 0.20%]	4.42e-09	12.74%	1.42%
All Index	68	8.01% [2.36%, 13.47%]	−2.68% [−3.51%, −0.01%]	0.11% [−0.27%, 0.37%]	8.17e-06	23.53%	0.00%
All Priors	144	3.33% [1.61%, 2.90%]	−0.53% [−0.37%, 0.03%]	0.03% [−0.15%, 0.16%]	1.46e-04	7.64%	2.08%
High Risk Exams	93	8.67% [3.89%, 13.46%]	−2.63% [−3.74%, 0.00%]	0.10% [−0.13%, 0.66%]	8.58e-07	29.03%	3.23%
All High-Risk Index	48	10.57% [4.64%, 15.49%]	−3.68% [−4.78%, 0.19%]	0.19% [−0.48%, 0.70%]	7.86e-05	33.33%	0.00%
All High-Risk Priors	45	6.65% [3.43%, 8.10%]	−1.51% [−2.24%, −0.04%]	0.00% [−0.06%, 0.66%]	0.003	24.44%	6.67%

*Low-risk predictions are defined as scores below the high-risk threshold.

[26/30] of localized predictions in cancer-affected breast for index exams, 65.5% [19/29] for priors). Priors without visible lesions showed a similar localization proportion in the cancer-affected (55% [11/20]) and contralateral breasts (45% [9/20]). Despite remaining high-risk upon case-occlusion, 10 of the 45 exams localized to the cancer-affected breast (22.2%) were considered a "lesion hit" because the centroid of the patch with the highest FOCUS Score fell within the (future) cancer region, suggesting that there were additional locations outside of the lesion region that also contribute to Mirai's high-risk estimate.

3.3.4 Qualitative Assessment

The 49 high-risk exams classified as either "lesion miss" or "contralateral breast localized" were reviewed to assess whether the patches with the highest FOCUS Score exhibited consistent patterns. Two common patterns emerged (Figure 3.8). The first pattern consisted of the presence of overlapping blood vessels. The second pattern reflected a sharp interface between high- and low-density regions (e.g., areas of fibroglandular-fat interface). Out of the

Table 3.3: Effect of case and control occlusion on Mirai's risk predictions on prior exam subsets.

Type of Prior Exam	**Exam Counts**	**Original Risk (mean, IQR)**	**Risk Reduction**		**p-value**	**Change to Low Risk***	
			Case	**Control**		**Case**	**Control**
All Priors	144	3.33% [1.61%, 2.90%]	−0.53% [−0.37%, 0.03%]	0.03% [−0.15%, 0.16%]	1.46e-04	7.64%	2.08%
≤ 2 years	56	3.60% [1.69%, 3.97%]	−0.96% [−0.51%, 0.00%]	0.09% [−0.12%, 0.16%]	0.001	7.14%	1.79%
>2 years	88	3.16% [1.59%, 2.72%]	−0.26% [−0.28%, 0.07%]	0.00% [−0.16%, 0.17%]	0.026	7.95%	2.27%
Lesion Visible	30	3.60% [1.83%, 3.95%]	−0.86% [−0.87%, −0.02%]	0.08% [−0.10%, 0.16%]	7.71e-04	3.33%	0.00%
Lesion Not Visible	114	3.26% [1.59%, 2.76%]	−0.45% [−0.30%, 0.04%]	0.02% [−0.16%, 0.16%]	0.013	8.77%	2.63%
All High-Risk Priors	45	6.65% [3.43%, 8.10%]	−1.51% [−2.24%, −0.04%]	0.00% [−0.06%, 0.66%]	0.003	24.44%	6.67%
High-Risk ≤ 2 years	21	6.60% [3.95%, 8.29%]	−2.36% [−3.61%, −0.12%]	0.30% [0.00%, 1.57%]	0.004	19.05%	4.76%
High-Risk > 2 years	24	6.69% [2.99%, 7.44%]	−0.76% [−1.16%, 0.13%]	−0.26% [−0.13%, 0.60%]	0.121	29.17%	8.33%
High-Risk with Lesion Visible	10	6.95% [4.20%, 8.00%]	−2.11% [−4.16%, −0.36%]	0.29% [0.02%, 1.02%]	0.037	10.00%	0.00%
High-Risk with Lesion Not Visible	35	6.57% [2.91%, 7.83%]	−1.33% [−1.28%, −0.02%]	−0.08% [−0.09%, 0.62%]	0.020	28.57%	8.57%

*Low-risk predictions are defined as scores below the high-risk threshold.

11 priors where case-occlusion shifted Mirai's prediction from high-risk to low-risk, 10 were deemed to have no visible lesion during annotation. Upon further visual inspection of the regions where cancer was subsequently detected, it was determined that subtle sub-clinical features of eventual cancer were present in 60.0% (6/10) of these exams (Figure 3.9).

Table 3.4: Categorization of mammography exams remaining high-risk after case-occlusion.

		Local Predictions		
		Case-Breast Localization		
Mammography Exam Type	**Global Predictions**	**Lesion Hit**	**Lesion Miss**	**Contralateral-Breast Localization**
All (n=66)	7 (10.6%)	10 (15.2%)	35 (53.0%)	14 (21.2%)
Index (n=32)	2 (6.3%)	6 (18.8%)	20 (62.5%)	4 (12.5%)
Priors (n=34)	5 (14.7%)	4 (11.8%)	15 (44.1%)	10 (29.4%)
Priors with lesion visible (n=9)	1 (11.1%)	3 (33.3%)	5 (55.5%)	0 (0.0%)
Priors with lesion not visible (n=25)	4 (16.0%)	1 (4%)	10 (40.0%)	10 (40.0%)

3.4 Discussion

We developed an analysis framework to quantitatively and qualitatively explain Mirai's breast cancer risk assessment. Masking cancer locations from Mirai's interpretation caused a significant reduction in its risk scores, even for prior exams without visible lesions, indicating that its assessments are significantly influenced by regions where breast cancers subsequently develop. Lesion-region masking shifted Mirai's assessment to low-risk for 29.0% of the exams predicted as high-risk (33.3% for index exams, 24.4% for priors), compared to 3.2% when masking anatomically matched regions in the contralateral breast. Even in exams that remained high-risk upon case-occlusion (the other 71.0%), our analysis suggests that Mirai's predictions tend to be strongly influenced by localized image regions. Nonetheless, the effects of cancer location masking were generally more pronounced in index exams, earlier priors, and priors with visible lesions. Thus, Mirai may use a range of strategies when estimating risk, including detecting existing cancers when present, and sometimes identifying subtle early signs of developing cancers before they are clinically visible, as our qualitative assessment suggests. Furthermore, the seven high-risk exams that

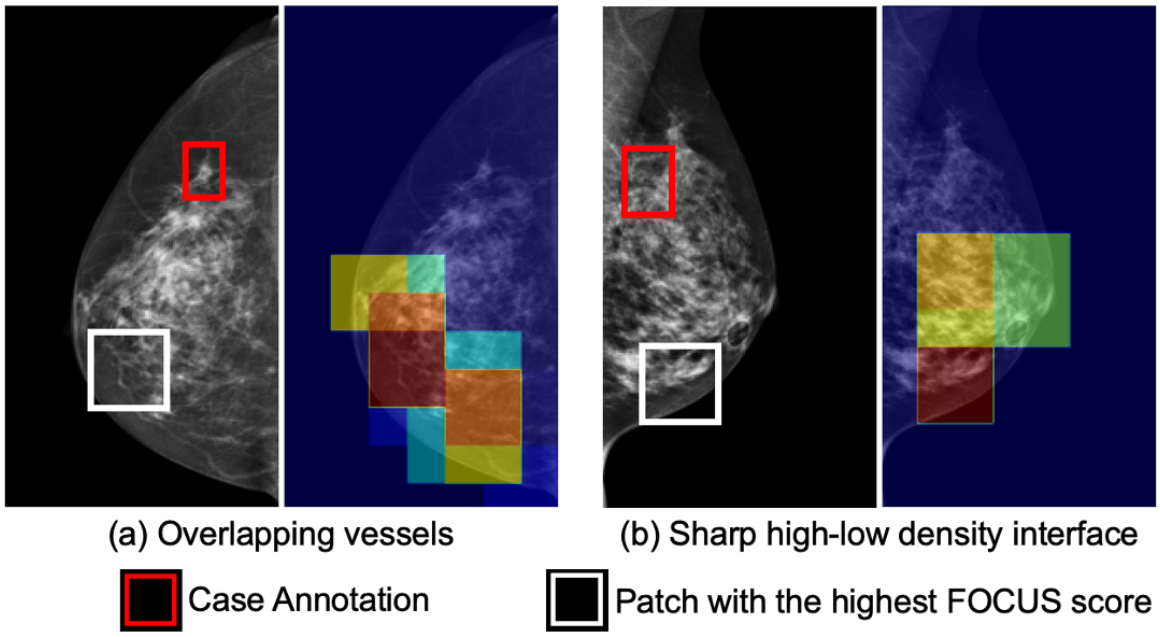
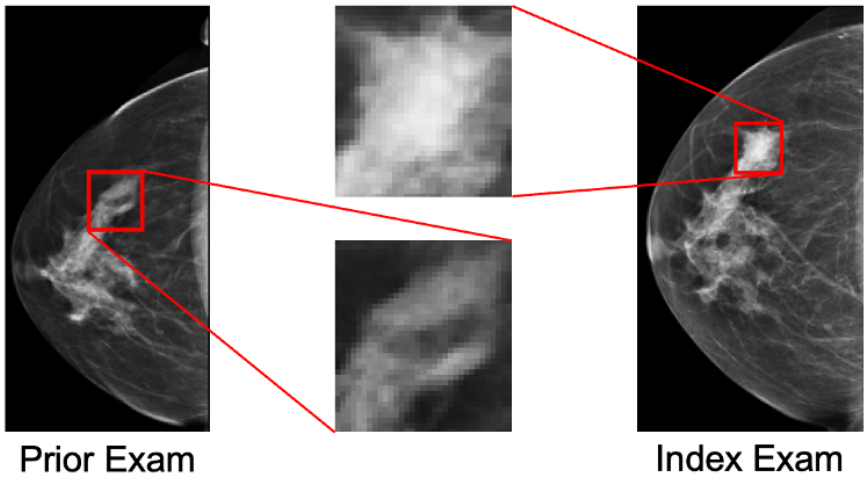


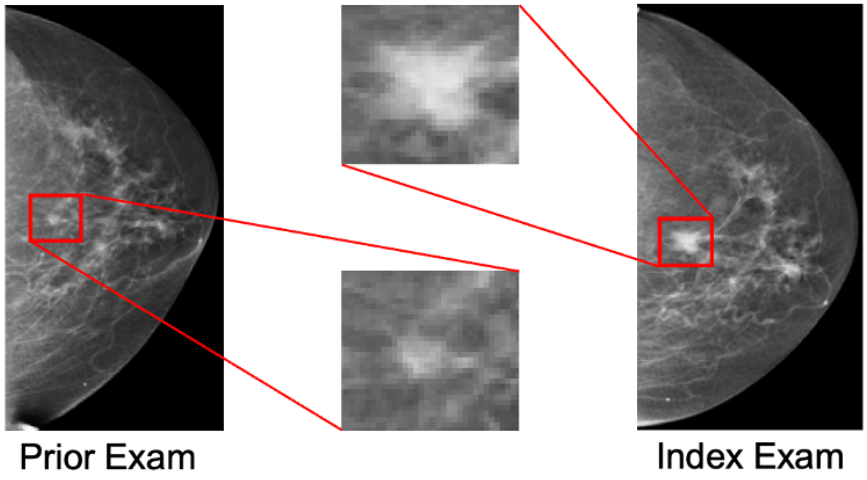
Figure 3.8: Qualitative assessment of Mirai's predictions localized to regions other than where cancer developed. (a) An index mammogram with a visible asymmetry. The patch with the highest FOCUS score highlights a region with overlapping vessels. (b) A five-year prior mammogram with visible calcifications. The patch with the highest FOCUS score highlights a sharp interface between high- and low-density regions.

exhibited global predictions were among women with dense breasts, suggesting that general parenchymal patterns can also drive Mirai's predictions. When predictions were localized to non-lesion regions, these areas often included overlapping vessels and sharp interfaces between high- and low-density tissue. Whether such features influence Mirai's score because they mimic lesions or reflect true risk factors remains unclear; we can only conclude that these regions contribute meaningfully to its risk predictions.

Our findings and FOCUS approach have several implications for clinical practice and early breast cancer detection. The often-localized nature of Mirai's predictions indicates opportunities to identify early signs of developing lesions and implement timely interventions. Our visualization technique, FOCUS Maps, could be used in practice to potentially help localize these regions of cancer development and explain Mirai's predictions for each



(a) Prior exam with no early findings visible.



(b) Prior exam with visible early findings.

Figure 3.9: Qualitative interpretation to assess whether Mirai detects the earliest signs of developing cancer. Analysis of prior exams where occlusion of the subsequent lesion region changed Mirai's interpretation from high-risk to low-risk: (a) Prior and index mammograms of a woman with no early sub-clinical findings in the prior exam, which subsequently showed a mass in the index exam. (b) Prior and index mammograms of a woman with early sub-clinical findings in the prior exam, which subsequently showed a focal asymmetry in the index exam.

exam more broadly. Ultimately, while further validation is necessary, these outputs could guide further clinical workup and inform the trustworthiness of Mirai's predictions.

By providing detailed localization analysis and categorization of Mirai's predictions, our work helps resolve outstanding questions from prior analyses. Omoleye et al. [42] assessed the influence of the cancer-affected breast on Mirai's risk estimates by mirroring either the cancer-affected or normal breast and found that cancer-affected breast mirroring resulted in similar performance, while mirroring the normal breast reduced performance to chance. Our results also suggest that the cancer-affected breast often drives Mirai's decisions, and, importantly, our approach provides a means to localize the score-driving regions of interest within the breast. Rather than explaining Mirai directly, Donnelly et al. [78] created a simplified version of Mirai, AsymMirai, which computes bilateral feature dissimilarity to generate risk estimates. While asymmetries may contribute to Mirai's predictions and breast cancer risk more generally, our results and those of Omoleye et al. [42] suggest that other features are also involved. In concurrent work, Wang et al. [79] examined individual features extracted by Mirai in its feature space and found that several correlate with the presence of calcifications and/or masses, and that the values of these features associate with Mirai's predictions. Our analysis also suggests that the presence of visible lesions influences Mirai's predictions, but that other, often localized, imaging features influence Mirai's predictions when lesions aren't present. Our work thus offers new insights through mechanistic interpretability that is faithful to Mirai's architecture. Additionally, while we focused on Mirai, our approach can be adapted to other imaging-based deep learning risk prediction models.

Our study has several limitations. While our visualizations demonstrate the potential to aid radiologists in interpreting Mirai's inferences, they require iteratively computing Mirai's predictions over different masked regions, which is computationally expensive. We mitigate this by utilizing the ClinValAI [43, 45] framework to implement an efficient cloud-based workflow, which can also be leveraged by others. Moreover, the visualizations currently rely on masking one patch at a time and do not consider potential non-linear effects when masking multiple patches. Additionally, our technique is based on Mirai's utilization of the Global Max Pooling layer. Though many deep learning models use this operation,

future work will involve extending this approach to other architectures. Our cohort had a limited sample size, and a single breast radiologist annotated the mammograms. However, the dataset was acquired from four imaging facilities that were not used during Mirai's development, and three breast radiologists independently verified the imaging findings, thus supporting the robustness of this external validation of Mirai's predictions.

Overall, our study demonstrates that while a range of features influences Mirai's predictions, its interpretations are often localized, particularly to the breast and location where cancer is subsequently detected. We also observed instances where Mirai detected early signs of developing cancer that were only detected by our radiologists in retrospect, knowing that cancer eventually developed in that location. These insights clarify the mechanisms of Mirai's risk predictions and highlight the potential utility of the FOCUS technique on a per-exam basis for targeting areas of high cancer risk with more focused and intensive imaging evaluation to facilitate earlier cancer detection.

Chapter 4

CONCLUSION

Artificial Intelligence (AI) algorithms, particularly deep learning models, show strong potential to assist both in mammography interpretation and future cancer risk prediction [22, 40, 84]. However, their integration into clinical workflows has been limited thus far, primarily due to a lack of computational tools to assess their performance and inadequate methods to explain their underlying decision-making process. In this dissertation, we have addressed these core challenges with novel techniques that can support health systems to evaluate the generalizability and interpretability of mammography-based deep learning algorithms. In this chapter, we summarize the key contributions, discuss the potential clinical impact, outline the limitations, and propose directions for future research.

Our informatics framework, ClinValAI [43, 45], as described in Chapter 2, addresses a need for a tool for health institutions to conduct external validation studies of AI algorithms while safeguarding patient imaging data privacy. By providing a customizable cloud-based infrastructure, our framework supports large-scale comparative evaluation of AI tools. We demonstrated ClinValAI's potential by validating multiple FDA-cleared commercial AI algorithms on a large dataset from various BCSC-affiliated breast cancer registries. This can help assess the potential benefits, such as assisting breast imaging radiologists in cancer detection, and the risks, such as false-positive assessments, of adopting FDA-approved commercial AI models in clinical settings, as shown in Figures 2.12 and 2.13.

By conducting rigorous assessments of deep learning algorithms for mammography using images from an institution's patient population, health organizations can gain a holistic understanding of their strengths (for example, Mirai's generalizability across subgroups defined by age, race, and breast density, as demonstrated in Figure 2.8) and weaknesses (for example, the reduction in Mirai's calibration metrics when exams with a breast cancer diagnosis within six months of screening are excluded, as shown in Table 2.3) for their specific

populations. Understanding the factors that influence AI's ability to guide clinicians in implementing appropriate interventions may help health systems improve the efficiency of mammography-based clinical workflows. For instance, external validation of AI algorithms can aid in determining exam-level thresholds to safely triage negative exams or function as standalone tools for mammography interpretation in specific subpopulations and scenarios. The ultimate goal of ClinValAI is to help ensure that only accurate, generalizable AI algorithms are adopted into clinical practice, thereby promoting improved and more equitable health outcomes.

However, our work has certain limitations. First, customizing ClinValAI's workflows requires expertise in cloud-based technologies and troubleshooting efforts. Second, large-scale external validation studies comparing the performance of multiple AI tools can incur considerable computational overhead. Thus, an important direction for future work would be to improve the reconfigurability and optimize the computational complexity of our validation pipelines. In addition, ClinValAI has not yet been used to support external validation of AI algorithms for medical imaging beyond mammography screening. In future, we may expand its capabilities to include other imaging modalities (such as ultrasound or MRI). Future work may also involve extending ClinValAI to support additional biomedical data modalities, including clinical text reports, omics data, and electronic health records (EHRs), as well as adapting the framework to enable validation of multimodal deep learning algorithms. Such extensions would require substantial customization of existing ClinValAI workflows and may require the development of new computational mechanisms to support validation of AI tools across a broader range of biomedical informatics applications.

Our work on the interpretability of Mirai's breast cancer risk prediction, as described in Chapter 3, can help radiologists, patients, primary care providers, and other screening stakeholders assess the trustworthiness of AI image-based risk estimation. While Mirai and other algorithms have the potential to streamline personalized screening strategies and risk reduction by identifying women at highest risk of developing future breast cancer, their black-box nature hinders widespread acceptance. Our explainability technique, FOCUS (Figure 3.2), helps analyze the contribution of different mammogram image regions to provide insight into Mirai's risk predictions by classifying Mirai's predictions into clinically meaningful se-

mantic categories (Figure 3.6). Our clinically interpretable visualizations, FOCUS Maps (Figure 3.5), also provide actionable insights by demonstrating Mirai's feature-importance distribution and identifing specific imaging regions that most strongly drive risk assessments.

While a range of factors drive Mirai's risk predictions, we found that they are strongly influenced by localized imaging features (Table 3.4), often corresponding to regions where cancers eventually develop (Tables 3.2, 3.3). This key finding suggests that the model can identify subtle, early signs of malignancy (Figure 3.9) that radiologists may not be able to detect. Our work also opens avenues for a broader understanding of the imaging features associated with breast cancer risk and development. Assessing whether or not the lesion versus non-lesion regions (Figure 3.8) influence Mirai's predictions could offer new insights into breast cancer radiomics and new avenues for directing more intense imaging work-up for patients at high risk.

Since our analysis is based solely on screening mammography exams from four University of Washington (UW) Medicine imaging facilities, an essential next step is to further validate our findings in larger samples across diverse cohorts. Moreover, while our explainability technique is currently specific to Mirai, it could be extended to evaluate other AI tools that receive FDA approval for clinical use. Insights into the inference mechanism of multiple deep learning algorithms may highlight similarities and differences in strategies involved in AI-based clinical decision support across products.

Overall, we established techniques for validating and explaining deep learning models, providing important insights prior to their widespread clinical adoption for mammography. More broadly, our work can advance research on the generalizability and interpretability of deep learning algorithms for medical imaging applications, supporting improved clinical workflows and promoting better, more equitable healthcare outcomes.

BIBLIOGRAPHY

[1] Arka Bhowmik and Sarah Eskreis-Winkler. Deep learning in breast imaging. *BJR Open*, 4(1):20210060, May 2022.

[2] Vanessa Buhrmester, David Münch, and Michael Arens. Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey, November 2019. arXiv:1911.12116 [cs].

[3] Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet (London, England)*, 380(9855):1778–1786, November 2012.

[4] Heidi D. Nelson, Kari Tyne, Arpana Naik, Christina Bougatsos, Benjamin K. Chan, Linda Humphrey, and U.S. Preventive Services Task Force. Screening for breast cancer: an update for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 151(10):727–737, W237–242, November 2009.

[5] László Tabár, Bedrich Vitak, Tony Hsiu-Hsi Chen, Amy Ming-Fang Yen, Anders Cohen, Tibor Tot, Sherry Yueh-Hsia Chiu, Sam Li-Sheng Chen, Jean Ching-Yuan Fann, Johan Rosell, Helena Fohlin, Robert A. Smith, and Stephen W. Duffy. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology*, 260(3):658–663, September 2011.

[6] S. Shapiro, W. Venet, P. Strax, L. Venet, and R. Roeser. Ten- to fourteen-year effect of screening on breast cancer mortality. *Journal of the National Cancer Institute*, 69(2):349–355, August 1982.

[7] Sylvia K. Plevritis, Diego Munoz, Allison W. Kurian, Natasha K. Stout, Oguzhan Alagoz, Aimee M. Near, Sandra J. Lee, Jeroen J. van den Broek, Xuelin Huang, Clyde B. Schechter, Brian L. Sprague, Juhee Song, Harry J. de Koning, Amy Trentham-

Dietz, Nicolien T. van Ravesteyn, Ronald Gangnon, Young Chandler, Yisheng Li, Cong Xu, Mehmet Ali Ergun, Hui Huang, Donald A. Berry, and Jeanne S. Mandelblatt. Association of Screening and Treatment With Breast Cancer Mortality by Molecular Subtype in US Women, 2000-2012. *JAMA*, 319(2):154–164, January 2018.

[8] Lydia E. Pace and Nancy L. Keating. A systematic assessment of benefits and risks to guide breast cancer screening decisions. *JAMA*, 311(13):1327–1335, April 2014.

[9] Mostafa Alabousi, Nanxi Zha, Jean-Paul Salameh, Lucy Samoilov, Anahita Dehmoobad Sharifabadi, Alex Pozdnyakov, Behnam Sadeghirad, Vivianne Freitas, Matthew D. F. McInnes, and Abdullah Alabousi. Digital breast tomosynthesis for breast cancer detection: a diagnostic test accuracy systematic review and meta-analysis. *European Radiology*, 30(4):2058–2071, April 2020.

[10] Gautam S. Muralidhar, Tamara Miner Haygood, Tanya W. Stephens, Gary J. Whitman, Alan C. Bovik, and Mia K. Markey. Computer-Aided Detection of Breast Cancer – Have All Bases Been Covered? *Breast Cancer : Basic and Clinical Research*, 2:5–9, May 2008.

[11] Stamatia V. Destounis, Patricia DiNitto, Wende Logan-Young, Ermelinda Bonaccio, Margarita L. Zuley, and Kathleen M. Willison. Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience. *Radiology*, 232(2):578–584, August 2004.

[12] R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. *Radiology*, 219(1):192–202, April 2001.

[13] Constance D. Lehman, Robert D. Wellman, Diana S. M. Buist, Karla Kerlikowske, Anna N. A. Tosteson, Diana L. Miglioretti, and Breast Cancer Surveillance Consortium. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA internal medicine*, 175(11):1828–1837, November 2015.

[14] Joshua J. Fenton, Stephen H. Taplin, Patricia A. Carney, Linn Abraham, Edward A. Sickles, Carl D'Orsi, Eric A. Berns, Gary Cutter, R. Edward Hendrick, William E. Barlow, and Joann G. Elmore. Influence of computer-aided detection on performance of screening mammography. *The New England Journal of Medicine*, 356(14):1399–1409, April 2007.

[15] Jiwoong J. Jeong, Brianna L. Vey, Ananth Bhimireddy, Thomas Kim, Thiago Santos, Ramon Correa, Raman Dutt, Marina Mosunjac, Gabriela Oprea-Ilies, Geoffrey Smith, Minjae Woo, Christopher R. McAdams, Mary S. Newell, Imon Banerjee, Judy Gichoya, and Hari Trivedi. The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.4 Million Screening and Diagnostic Mammographic Images. *Radiology: Artificial Intelligence*, 5(1):e220047, January 2023. Publisher: Radiological Society of North America.

[16] CBIS-DDSM.

[17] Hieu T. Nguyen, Ha Q. Nguyen, Hieu H. Pham, Khanh Lam, Linh T. Le, Minh Dao, and Van Vu. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, May 2023. Publisher: Nature Publishing Group.

[18] Gaurav Bhole, S. Suba, and Nita Parekh. Mammo-Bench: A Large-scale Benchmark Dataset of Mammography Images, February 2025. Pages: 2025.01.31.25321510.

[19] Keiron O'Shea and Ryan Nash. An Introduction to Convolutional Neural Networks, December 2015. arXiv:1511.08458 [cs].

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. Publisher: Nature Publishing Group.

[22] William Lotter, Abdul Rahman Diab, Bryan Haslam, Jiye G. Kim, Giorgia Grisot, Eric Wu, Kevin Wu, Jorge Onieva Onieva, Yun Boyer, Jerrold L. Boxerman, Meiyun Wang, Mack Bandler, Gopal R. Vijayaraghavan, and A. Gregory Sorensen. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature Medicine*, 27(2):244–249, February 2021. Publisher: Nature Publishing Group.

[23] Sarah E. Hickman, Nicholas R. Payne, Richard T. Black, Yuan Huang, Andrew N. Priest, Sue Hudson, Bahman Kasmai, Arne Juette, Muzna Nanaa, and Fiona J. Gilbert. Deep Learning Algorithms for Breast Cancer Detection in a UK Screening Cohort: As Stand-alone Readers and Combined with Human Readers. *Radiology*, 313(2):e233147, November 2024. Publisher: Radiological Society of North America.

[24] Naresh Khuriwal and Nidhi Mishra. Breast Cancer Diagnosis Using Deep Learning Algorithm. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 98–103, October 2018.

[25] Karin Dembrower, Yue Liu, Hossein Azizpour, Martin Eklund, Kevin Smith, Peter Lindholm, and Fredrik Strand. Comparison of a Deep Learning Risk Score and Standard Mammographic Density Score for Breast Cancer Risk Prediction. *Radiology*, 294(2):265–272, February 2020. Publisher: Radiological Society of North America.

[26] Adam Yala, Peter G. Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13(578):eaba4373, January 2021. Publisher: American Association for the Advancement of Science.

[27] Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, and Hua Xu. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association : JAMIA*, 27(3):457–470, December 2019.

[28] Jonathan Tyrer, Stephen W. Duffy, and Jack Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in Medicine*, 23(7):1111–1130, April 2004.

[29] Breast Cancer Risk Assessment Tool: Online Calculator (The Gail Model).

[30] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24):1879–1886, December 1989.

[31] Thomas Schaffter, Diana S. M. Buist, Christoph I. Lee, Yaroslav Nikulin, Dezső Ribli, Yuanfang Guan, William Lotter, Zequn Jie, Hao Du, Sijia Wang, Jiashi Feng, Mengling Feng, Hyo-Eun Kim, Francisco Albiol, Alberto Albiol, Stephen Morrell, Zbigniew Wojna, Mehmet Eren Ahsen, Umar Asif, Antonio Jimeno Yepes, Shivanthan Yohanandan, Simona Rabinovici-Cohen, Darvin Yi, Bruce Hoff, Thomas Yu, Elias Chaibub Neto, Daniel L. Rubin, Peter Lindholm, Laurie R. Margolies, Russell Bailey McBride, Joseph H. Rothstein, Weiva Sieh, Rami Ben-Ari, Stefan Harrer, Andrew Trister, Stephen Friend, Thea Norman, Berkman Sahiner, Fredrik Strand, Justin Guinney, Gustavo Stolovitzky, and and the DM DREAM Consortium. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Network Open*, 3(3):e200265, March 2020.

[32] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, January 2020. Publisher: Nature Publishing Group.

[33] Anna W. Anderson, M. Luke Marinovich, Nehmat Houssami, Kathryn P. Lowry, Joann G. Elmore, Diana S. M. Buist, Solveig Hofvind, and Christoph I. Lee. Independent External Validation of Artificial Intelligence Algorithms for Automated Interpretation of Screening Mammography: A Systematic Review. *Journal of the American College of Radiology: JACR*, 19(2 Pt A):259–273, February 2022.

[34] Alice C. Yu, Bahram Mohajer, and John Eng. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. *Radiology. Artificial Intelligence*, 4(3):e210064, May 2022.

[35] William Hsu, Daniel S. Hippe, Noor Nakhaei, Pin-Chieh Wang, Bing Zhu, Nathan Siu, Mehmet Eren Ahsen, William Lotter, A. Gregory Sorensen, Arash Naeim, Diana S. M. Buist, Thomas Schaffter, Justin Guinney, Joann G. Elmore, and Christoph I. Lee. External Validation of an Ensemble Model for Automated Mammography Interpretation by Artificial Intelligence. *JAMA Network Open*, 5(11):e2242343, November 2022.

[36] David C. Classen, Christopher Longhurst, and Eric J. Thomas. Bending the patient safety curve: how much can AI help? *npj Digital Medicine*, 6(1):2, January 2023. Publisher: Nature Publishing Group.

[37] Farhad Maleki, Katie Ovens, Rajiv Gupta, Caroline Reinhold, Alan Spatz, and Reza Forghani. Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. *Radiology: Artificial Intelligence*, 5(1):e220028, November 2022.

[38] Yan Gao, Teena Sharma, and Yan Cui. Addressing the Challenge of Biomedical Data Inequality: An Artificial Intelligence Perspective. *Annual review of biomedical data science*, 6:153–171, August 2023.

[39] Zhan Xu, David E. Rauch, Rania M. Mohamed, Sanaz Pashapoor, Zijian Zhou, Bikash Panthi, Jong Bum Son, Ken-Pin Hwang, Benjamin C. Musall, Beatriz E. Adrada, Rosalind P. Candelaria, Jessica W. T. Leung, Huong T. C. Le-Petross, Deanna L. Lane, Frances Perez, Jason White, Alyson Clayborn, Brandy Reed, Huiqin Chen,

Jia Sun, Peng Wei, Alastair Thompson, Anil Korkut, Lei Huo, Kelly K. Hunt, Jennifer K. Litton, Vicente Valero, Debu Tripathy, Wei Yang, Clinton Yam, and Jingfei Ma. Deep Learning for Fully Automatic Tumor Segmentation on Serially Acquired Dynamic Contrast-Enhanced MRI Images of Triple-Negative Breast Cancer. *Cancers*, 15(19):4829, October 2023.

[40] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology*, 292(1):60–66, July 2019. Publisher: Radiological Society of North America.

[41] Adam Yala, Peter G. Mikhael, Fredrik Strand, Gigin Lin, Siddharth Satuluru, Thomas Kim, Imon Banerjee, Judy Gichoya, Hari Trivedi, Constance D. Lehman, Kevin Hughes, David J. Sheedy, Lisa M. Matthis, Bipin Karunakaran, Karen E. Hegarty, Silvia Sabino, Thiago B. Silva, Maria C. Evangelista, Renato F. Caron, Bruno Souza, Edmundo C. Mauad, Tal Patalon, Sharon Handelman-Gotlib, Michal Guindy, and Regina Barzilay. Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. *Journal of Clinical Oncology*, 40(16):1732–1740, June 2022. Publisher: Wolters Kluwer.

[42] Olasubomi J. Omoleye, Anna E. Woodard, Frederick M. Howard, Fangyuan Zhao, Toshio F. Yoshimatsu, Yonglan Zheng, Alexander T. Pearson, Maksim Levental, Benjamin S. Aribisala, Kirti Kulkarni, Gregory S. Karczmar, Olufunmilayo I. Olopade, Hiroyuki Abe, and Dezheng Huo. External Evaluation of a Mammography-based Deep Learning Model for Predicting Breast Cancer in an Ethnically Diverse Population. *Radiology: Artificial Intelligence*, 5(6):e220299, November 2023. Publisher: Radiological Society of North America.

[43] Ojas A. Ramwala, Kathryn P. Lowry, Daniel S. Hippe, Matthew P. N. Unrath, Matthew J. Nyflot, Sean D. Mooney, and Christoph I. Lee. ClinValAI: A framework for developing Cloud-based infrastructures for the External Clinical Validation of AI

in Medical Imaging. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 30:215–228, 2025.

[44] Celeste Damiani, Grigorios Kalliatakis, Muthyala Sreenivas, Miaad Al-Attar, Janice Rose, Clare Pudney, Emily F. Lane, Jack Cuzick, Giovanni Montana, and Adam R. Brentnall. Evaluation of an AI Model to Assess Future Breast Cancer Risk. *Radiology*, 307(5):e222679, June 2023. Publisher: Radiological Society of North America.

[45] Ojas A. Ramwala, Kathryn P. Lowry, Nathan M. Cross, William Hsu, Christopher C. Austin, Sean D. Mooney, and Christoph I. Lee. Establishing a Validation Infrastructure for Imaging-Based Artificial Intelligence Algorithms Before Clinical Implementation. *Journal of the American College of Radiology*, 21(10):1569–1574, October 2024.

[46] Michael Fatemi, Eric Feng, Cyril Sharma, Zarif Azher, Tarushii Goel, Ojas Ramwala, Scott M. Palisoul, Rachael E. Barney, Laurent Perreard, Fred W. Kolling, Lucas A. Salas, Brock C. Christensen, Gregory J. Tsongalis, Louis J. Vaickus, and Joshua J. Levy. Inferring spatial transcriptomics markers from whole slide images to characterize metastasis-related spatial heterogeneity of colorectal tumors: A pilot study. *Journal of Pathology Informatics*, 14:100308, January 2023.

[47] Cody M. Schopf, Ojas A. Ramwala, Kathryn P. Lowry, Solveig Hofvind, M. Luke Marinovich, Nehmat Houssami, Joann G. Elmore, Brian N. Dontchos, Janie M. Lee, and Christoph I. Lee. Artificial Intelligence-Driven Mammography-Based Future Breast Cancer Risk Prediction: A Systematic Review. *Journal of the American College of Radiology*, 2023. Publisher: Elsevier.

[48] Ojas A. Ramwala, Smeet A. Dhakecha, Antriksh Ganjoo, Divyanshu Visiya, and Jignesh N. Sarvaiya. Leveraging Adversarial Training for Efficient Retinal Vessel Segmentation. In *2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6, July 2021.

[49] COVID-19 Diagnosis from Chest Radiography Images using Deep Residual Network | IEEE Conference Publication | IEEE Xplore.

[50] Himansh Mulchandani, Poojan Dalal, Ojas A. Ramwala, Parima Parikh, Upena Dalal, Mita Paunwala, and Chirag Paunwala. Tonsillitis based Early Diagnosis of COVID-19 for Mass-Screening using One-Shot Learning Framework. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–6, December 2020. ISSN: 2325-9418.

[51] Novel Multi-Modal Throat Inflammation and Chest Radiography based Early-Diagnosis and Mass-Screening of COVID-19. *The Open Biomedical Engineering Journal*, 15:226–234, January 2021.

[52] P. Dalal, M. Himansh, O. Ramwala, P. Parikh, U. Dalal, M. Paunwala, and C. Paunwala. Throat Inflammation Based Mass Screening of Covid-19 on Embedded Platform. In Kanubhai K. Patel, Deepak Garg, Atul Patel, and Pawan Lingras, editors, *Soft Computing and its Engineering Applications*, pages 277–288, Singapore, 2021. Springer.

[53] Joshua Levy, Yunrui Lu, Marietta Montivero, Ojas Ramwala, Jason McFadden, Carly Miles, Adam Gilbert Diamond, Ramya Reddy, Ram Reddy, Taylor Hudson, Zarif Azher, Akash Pamal, Sameer Gabbita, Tess Cronin, Abdol Aziz Ould Ismail, Tarushii Goel, Sanjay Jacob, Anish Suvarna, Taein Kim, Edward Zhang, Neha Reddy, Sumanth Ratna, Jason Zavras, and Louis Vaickus. Artificial Intelligence, Bioinformatics, and Pathology: Emerging Trends Part II—Current Applications in Anatomic and Molecular Pathology. *Advances in Molecular Pathology*, 5(1):e25–e52, November 2022. Publisher: Elsevier.

[54] Joshua Levy, Yunrui Lu, Marietta Montivero, Ojas Ramwala, Jason McFadden, Carly Miles, Adam Gilbert Diamond, Ramya Reddy, Ram Reddy, Taylor Hudson, Zarif Azher, Akash Pamal, Sameer Gabbita, Tess Cronin, Abdol Aziz Ould Ismail, Tarushii Goel, Sanjay Jacob, Anish Suvarna, Sumanth Ratna, Jason Zavras, and Louis Vaickus. Artificial Intelligence, Bioinformatics, and Pathology: Emerging Trends Part I—an Introduction to Machine Learning Technologies. *Advances in Molecular Pathology*, 5(1):e1–e24, November 2022. Publisher: Elsevier.

[55] Center for Devices and Radiological Health. Artificial Intelligence-Enabled Medical Devices. *FDA*, December 2025. Publisher: FDA.

[56] Florence X. Doo, Pranav Kulkarni, Eliot L. Siegel, Michael Toland, Paul H. Yi, Ruth C. Carlos, and Vishwa S. Parekh. Economic and Environmental Costs of Cloud Technologies for Medical Imaging and Radiology Artificial Intelligence. *Journal of the American College of Radiology: JACR*, 21(2):248–256, February 2024.

[57] Andrea Alonso and Jeffrey J. Siracuse. Protecting patient safety and privacy in the era of artificial intelligence. *Seminars in Vascular Surgery*, 36(3):426–429, September 2023.

[58] Timothy Bergquist, Yao Yan, Thomas Schaffter, Thomas Yu, Vikas Pejaver, Noah Hammarlund, Justin Prosser, Justin Guinney, and Sean Mooney. Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction. *Journal of the American Medical Informatics Association : JAMIA*, 27(9):1393–1400, July 2020.

[59] Yao Yan. *A Model-to-data Approach for Building Accurate Machine Learning Algorithms on EHR Data*. Thesis, 2022. Accepted: 2022-09-23T20:42:01Z.

[60] Anton S. Becker, Magda Marcon, Soleen Ghafoor, Moritz C. Wurnig, Thomas Frauenfelder, and Andreas Boss. Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. *Investigative Radiology*, 52(7):434–440, July 2017.

[61] Stephen J. Mooney and Vikas Pejaver. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annual Review of Public Health*, 39(Volume 39, 2018):95–112, April 2018. Publisher: Annual Reviews.

[62] Justin Guinney and Julio Saez-Rodriguez. Alternative models for sharing confidential biomedical data. *Nature Biotechnology*, 36(5):391–392, May 2018. Publisher: Nature Publishing Group.

[63] Kate Voss, Jeff Gentry, and Geraldine Van Der Auwera. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. 2017. Publisher: F1000Research.

[64] miniwdl — miniwdl documentation.

[65] Acr. *2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. American College of Radiology, January 2014. Google-Books-ID: nhWSjwEACAAJ.

[66] Francis L. Huang. Using Cluster Bootstrapping to Analyze Nested Data With a Few Clusters. *Educational and Psychological Measurement*, 78(2):297–318, April 2018.

[67] Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, May 2011.

[68] Cynthia S. Crowson, Elizabeth J. Atkinson, and Terry M. Therneau. Assessing Calibration of Prognostic Risk Scores. *Statistical methods in medical research*, 25(4):1692–1706, August 2016.

[69] Breast Cancer Surveillance Consortium (BCSC).

[70] Orthanc - DICOM Server.

[71] Dooman Arefan, Aly A. Mohamed, Wendie A. Berg, Margarita L. Zuley, Jules H. Sumkin, and Shandong Wu. Deep learning modeling using normal mammograms for predicting breast cancer risk. *Medical physics*, 47(1):110–118, January 2020.

[72] Saba Dadsetan, Dooman Arefan, Wendie A. Berg, Margarita L. Zuley, Jules H. Sumkin, and Shandong Wu. Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. *Pattern recognition*, 132:108919, December 2022.

[73] Vignesh A. Arasu, Laurel A. Habel, Ninah S. Achacoso, Diana S. M. Buist, Jason B. Cord, Laura J. Esserman, Nola M. Hylton, M. Maria Glymour, John Kornak, Lawrence H. Kushi, Donald A. Lewis, Vincent X. Liu, Caitlin M. Lydon, Diana L. Miglioretti, Daniel A. Navarro, Albert Pu, Li Shen, Weiva Sieh, Hyo-Chun Yoon, and

Catherine Lee. Comparison of Mammography AI Algorithms with a Clinical Risk Model for 5-year Breast Cancer Risk Prediction: An Observational Study. *Radiology*, 307(5):e222733, June 2023. Publisher: Radiological Society of North America.

[74] Constance D Lehman, Sarah Mercaldo, Leslie R Lamb, Tari A King, Leif W Ellisen, Michelle Specht, and Rulla M Tamimi. Deep Learning vs Traditional Breast Cancer Risk Models to Support Risk-Based Mammography Screening. *JNCI: Journal of the National Cancer Institute*, 114(10):1355–1363, October 2022.

[75] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, October 2017. ISSN: 2380-7504.

[76] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[77] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, Julius Adebayo, Matthew D. Li, and Jayashree Kalpathy-Cramer. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, November 2021. Publisher: Radiological Society of North America.

[78] Jon Donnelly, Luke Moffett, Alina Jade Barnett, Hari Trivedi, Fides Schwartz, Joseph Lo, and Cynthia Rudin. AsymMirai: Interpretable Mammography-based Deep Learning Model for 1–5-year Breast Cancer Risk Prediction. *Radiology*, 310(3):e232780, March 2024. Publisher: Radiological Society of North America.

[79] Yao-Kuan Wang, Zan Klanecek, Tobias Wagner, Lesley Cockmartin, Nicholas Marshall, Andrej Studen, Robert Jeraj, and Hilde Bosmans. Using Explainable AI to Characterize Features in the Mirai Mammographic Breast Cancer Risk Prediction Model. *Radi-*

ology: Artificial Intelligence, page e240417, September 2025. Publisher: Radiological Society of North America.

[80] Abhishek Dutta and Andrew Zisserman. The VIA Annotation Software for Images, Audio and Video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 2276–2279, New York, NY, USA, October 2019. Association for Computing Machinery.

[81] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[83] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945. Publisher: [International Biometric Society, Wiley].

[84] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE transactions on medical imaging*, 39(4):1184–1194, April 2020.