DISSERTATION

At

Mahajan Imaging Research and Development Pvt. Ltd, New Delhi

Report on

Validation of AI models on Medical Images

By

Dr. Bhanushree Bahl

PG/19-21/019

Health IT management

Under the guidance of: Dr. Pankaj Talreja

POST GRADUATE DIPLOMA IN HOSPITAL AND HEALTH MANANGEMENT

2018-20

International Institute of Health Management Research
New Delhi

(Completion of Dissertation from respective organization)

The certificate is awarded to

Name: Dr. Bhanushree Bahl

In recognition of having successfully completed her
Internship in the department of

Title: Caring Analytics Platform (CARPL)

And has successfully completed her Project on

Validation of AI Models on Medical Images

Date: 1st March 2021-1st June 2021

Organization: Mahajan Imaging Research and Development Pvt. Ltd.

She comes across as a committed, sincere & diligent person
who has a strong drive & zeal for learning.

We wish her all the best for future endeavors.

Training & Development
Dr. Vasantha Kumar Venugopal, MD
(Medical Imaging Lead, CARING)

Zonal Head-Human Resources
Mrs. Shalu Mamgain

TO WHOMSOEVER IT MAY CONCERN

This is to certify that **Dr. Bhanushree Bahl** student of PGDM (Hospital & Health Management) from International Institute of Health Management Research, New Delhi has undergone internship training at **Mahajan Imaging Research and Development Pvt. Ltd.** from 1st May 2021 to 1st June 2021.

The Candidate has successfully carried out the study designated to her during internship training and her approach to the study has been sincere, scientific and analytical.

The Internship is in fulfillment of the course requirements.

I wish her all success in all her future endeavors.

Dr. Pankaj Talreja
Professor,
IIHMR, New Delhi

Mentor
IIHMR, New Delhi

Certificate of Approval

The following dissertation titled **“Validation of AI models on Medical Images”** at **“Mahajan Imaging Research and Development Pvt. Ltd.”** is hereby approved as a certified study in management carried out and presented in a manner satisfactorily to warrant its acceptance as a prerequisite for the award of **Post Graduate Diploma in Health and Hospital Management** for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the dissertation only for the purpose it is submitted.

Dissertation Examination Committee for evaluation of dissertation.

Name	Signature
____________________	____________________
____________________	____________________
____________________	____________________

INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT RESEARCH, NEW DELHI

CERTIFICATE BY SCHOLAR

This is to certify that the dissertation titled **Validation of AI models on Medical Images** and submitted by **Dr. Bhanushree Bahl** Enrollment No. **PG/19-21/019** under the supervision of **Dr. Pankaj Talreja** for award of PGDM (Hospital & Health Management) of the Institute carried out during the period from **1st May 2021 to 1st June 2021** embodies my original work and has not formed the basis for the award of any degree, diploma associate ship, fellowship, titles in this or any other Institute or other similar institution of higher learning.

Signature

FEEDBACK FORM

Name of the Student: Dr. Bhanushree Bahl

Dissertation Organization: Mahajan Imaging Research and Development Pvt. Ltd.

Area of Dissertation: Caring Analytics Platform (CARPL)

Attendance: 100%

Objectives achieved:

1. Validation of AI models
2. CARPL Use cases and Workflow Diagrams
3. CARPL Demo to the clients
4. Instruction Videos for the platform (CARPL) for clients

Deliverables:

1. Using CARPL validated third party AI models on medical images
2. Using E-Draw as a tool, made workflow diagrams and delivered use cases for 4 modules of the platform
3. Gave Demo for different modules of the platform to the clients
4. Made Instruction videos of the platform using video editor for clients

Strengths:

1. Time Management, Dedication and Hard work- Bhanushee was assigned multiple tasks which was new for her but still she managed to complete them successfully
2. Teamwork and Communication – She was working well with her peers in the group project, and also communicating her understanding, progress and queries to us in a timely and descriptive manner.
3. Quick/Fast Learner – She got the hang of our business very quickly and picked up on the requirements of her tasks pretty soon as well

Suggestions for Improvement:

1. She need to work on her excel skills and can take up some data analytics courses.

Suggestions for Institute (course curriculum, industry interaction, placement, alumni):

1. Institute should provide thorough knowledge of the roles students are applying for.
2. Basic computer skills should be considered as an altogether module for students.

Signature of the Officer-in-Charge/ Organization Mentor (Dissertation)
Dr. Vasantha Kumar Venugopal, MD
(Medical Imaging Lead, CARING)

Date: 31st May, 2021
Place: New Delhi

ACKNOWLEDGEMENTS

I would like to express my special thanks to **Dr. Anandhi Ramachandran (Associate Professor*)*** and IIHMR, Delhi for giving me the golden opportunity to pursue my dissertation at CARING, MAHAJAN IMAGING PVT. LTD, New Delhi.

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of my dissertation at CARING, MAHAJAN IMAGING PVT. LTD, New Delhi.
I am highly thankful to **Dr. Vidur Mahajan (Head R&D)** CARING, MAHAJAN IMAGING PVT. LTD, for granting me the Opportunity for training and, for being helpful, supportive and receptive. His broad and profound knowledge, critical thinking has given me constant encouragement to achieve the task allotted and perform better.

.

Also, my successful completion of the tasks would not have been possible without the helping deeds from the entire department of CARING and especially **Dr. Vasantha Kumar Venugopal and Mrs. Kabita Dash** as they were always so cooperative and encouraging. Without their co-operation and warm attitude this project would not have been possible.

Lastly, I would also like to thank my parents and friends who helped me a lot in being mentally strong and helping me complete my report directly or indirectly.

Special Regards to my mentor **Dr. Pankaj Talreja** (Professor, IIHMR Delhi) for his constant supervision and support in completing the projects.

Contents

ORGANIZATION PROFILE

About CARING:

CARING, short for Centre for Advanced Research in Imaging, Neuroscience and Genomics, is Mahajan Imaging's newly christened research & development wing focused on performing cutting-edge scientific and clinical research and helping radiology and genomics companies develop world-class clinically relevant products.

CARING currently works with 15+ collaborators including academia, start-ups and industry and is open to working with imaging researchers & engineers, neuroscientists and genomic medicine experts on developing insights and products for a better tomorrow.

The CARING Analytics Platform – CARPL – is the world's first and only end-to-end development, testing and deployment platform for medical imaging AI applications. CARPL is used by the world's leading data scientists, startups, medical imaging companies, academic centers and hospitals to help ensure safe and effortless deployment of AI. CARPL comprises a data management & search platform, an annotation platform, a pre-deployment testing platform and an algorithm deployment platform. You can learn more about CARPL from this presentation at the IIT Mumbai Tech Fest, 2020.

CARPL is built by CARING – the Centre for Advanced Research in Imaging, Neurosciences & Genomics – a group of clinicians, engineers and scientists who are frustrated with the lack of adoption of AI and digital solutions in healthcare. CARING is the product group at Mahajan Imaging, India's leading and most advanced diagnostics service provider which caters to more than 500,000 patients per year. We are currently bootstrapped and profitable!

We publish prolifically and have more than 100 papers presented at leading conferences across the world, and more than 10 journal papers, including the first paper on AI in the Lancet. We work hard, party harder and love being the best at what we do ☺ and are now growing!

1. PREFACE

1.1 Abstract

[Keywords- AI, Validation, CARPL, FN, FP, Dynamic Thresholds]

With increase in magnification of artificial intelligence (AI) and other technologies using machine learning, many models have been developed to provide intelligent decisions based on their input to acquire features and functionalities that could detect diseases, yield prediction and provide quick decision making on diagnosis. With regard to diagnostic and predictive analysis, usage of AI is a point of extreme interest. Quality of AI models are of great concern here, to implement an AI model into clinical environment, validation of the algorithm is of utmost priority. This research focuses on various performance of AI models on medical images, various methods of validation, issues that were faced, methods on improving the algorithm. Three external validation were conducted on 3 third party AI models and there performance was evaluated using CARPL platform as statistical tool. Other than 3 case studies, 6 articles were reviewed for the study which included one trained and tested AI algorithm for COVID-19. The results were impressive and concluded that the AI model could be used for triaging patients with COVID-19. The study concludes that validation for any AI algorithm is imperative before deploying it on any clinical practice. Considerable amount clinical trials should be performed using external data to evaluate the performance of the model. Lastly, to improve the algorithm examination of FN and FP and dynamic thresholds is essential to yield positive results and for patient benefit.

1.2 ABBREVATIONS

AI	Artificial Intelligence
ROC	Receiver Operating Characteristics
AUC	Area Under Curve
UI	User Interface
NPV	Negative Predictive Value
PPV	Positive Predictive Value
MCC	Matthews Correlation Coefficient
FN	False Negatives
FP	False Positives

2. DISSERTATION REPORT

2.1 Introduction

The growth of AI in the medicine world has been in talk for a while now. More the advancements in deep neural technologies in predicting and diagnosing medical images more it is becoming of great interest. For example photographs of retina, skin lesions, pathological or radiological images, all of these have great potential usefulness due to the advancements in the AI technologies. Now before approval of any AI algorithm into the clinical practice, it is essential that a thorough testing, performance and utility has been achieved.
For every other medical devices before adopting it into the clinical domain, a complete validation is performed similarly before adoption of any AI technology tool a thorough validation of any AI algorithm is important. A comprehensive validation of AI algorithm will ensure patient benefit and safety, and will also bypass unintended harms.
There are three ways of measuring the validation of an AI tool into clinical practice: its Diagnostic performance, response to patient outcome and communal efficacy. For quality assessment real-time clinical validation is necessary of any AI algorithm on medical images which uses deep neural technology. For appropriate confirmation external validation is recommended which requires sufficient assessed datasets. Now these dataset can be gathered either from new patients or from associations which can provide training data of the target patients where AI technology can be applied. Advantage of using external data is that it will certify algorithm's ability to generalize across the variables in different clinical systems.
Algorithms that analyses medical images, requires an extensive amount of data for training the model and annotating the images, which becomes a challenge to surpass for most of AI developing companies. Since algorithms truly depend on their training data there is an absolute risk that it may not perform strong in the clinical practice or real-time setting or it is not certain that if the algorithm has given accurate outputs at one institution it will perform the same at other.

Statistical techniques used for validation of AI models:

a. **Discrimination Performance**- This refers to when a diagnostic test or algorithm gives outputs in a binary classification. These are frequently measured in terms of sensitivity, which refers to patients who test positive actually have the abnormality and specificity, which refers to patients who test negative and do not have the disease. For determining the discrimination performance of an algorithms/model ROC analysis is an effective statistical tool to measure. At different thresholds, various pairs of sensitivity and specificity values are gained. As threshold decreases for a disease there is an increase in sensitivity whereas there is a decrease specificity and vice-versa. By using different points one can plot a graph between sensitivity (True positives) as y axis and 1-specificity (False positives) on x axis. AUC (area under the curve) which a common measure for an ROC curve which can be interpreted as the average value of positive cases for all possible values of negative cases or average value of negative cases for all possible values of positives. This ranges from 0-1, if AUC value is closer to 1, better is the performance of the model.

b. **Calibration Performance**- Instead of breaking the result into binary classification, this refers to giving the output as how alike are the probabilities of the predicted model to the actual probabilities. The calibration graph is plotted between predicted probabilities on y axis and real/actual probabilities on x axis.

c. **Use of internal versus external data**- When a particular set data is used to train or develop the model to determine its performance it refers to as internal validation, and when a different dataset is used to assess the performance of the model is known as external validation. Hence, extraneous validation is critical in order to verify the diagnostic ability of a model for predictions.

d. **Patient outcome verification**- The end goal for any advancement in medical world is patient safety and benefit. Similarly, for any deep learning tool patient benefit is the ultimate goal, an alternative where one could determine whether the model would benefit the patients or not if used in real-world should be considered.

AI Software Testing: The main concern here is validating the AI software functions, behaviors and output. This process includes planning, model testing, and generation of testing case, its implementation and determination. Few techniques that can be adopted are:

1. Decision table testing design technique- Here one evaluates various combination of inputs associated with their outputs and rules the system.
2. Black-box testing- Here testing is used to evaluate the user prerequisites for example, to discover the failures in the following-erroneous functions, UI failures, performance failures etc.

Common Validation Methodology for AI software:

a. Classification based AI software testing- This is performed for sufficient testing coverage for various input data classes and their output classes, for contexts and conditions.
b. Model based AI software testing- This refers to choosing developing and data models to be detectable and can be tested so as to ease the AI system evaluation and operating in training and testing the data.
c. Learning based AI software testing using the crowd-sourced approach- Here in a service platform, crowd based testers are used to learn from machine learning models and different approaches.

Dataset assessment for AI based system:

a. Raw data quality checking- This process refers to quality assessment of collected organic data for example, camera generated images and videos. The fundamental goal is to clean, monitor the quality and evaluation of the raw data that has been collected.

b. Training data quality validation- This process involves quality assessment of the training data or annotated datasets. Its aim is to enhance the generation of the training data on a deep learning model in order to enhance the aspect of the AI system associated with it.

c. Test data quality validation- This refers to assessing the test data quality based on the validation results of an application. The focus should be on detection, bug improvement, training quality coverage for AI models.

Testing framework and its quality assessment for AI system-

a. Correctness: This refers to Boolean items and if their results are true, for example-gender, age group, diseased or not.
b. Accuracy: This refers to accuracy of results with various items like age, gender. These can be measured using confidence level, absolute/relative mean.
c. System stability: This refers to when a system is tested once or twice or thrice the stability of the system should remain same.
d. Timeliness: This process is related to implications related to time, for example- training time, classifying time, recognition time.
e. System robustness: This process implies to robustness of the system for example when performing special operations on an image does the system recognizes the image well.
f. Image quality: This process puts a check if the system can deal with change in the quality of the images.

DeepCOVID-XR- An algorithm model that was developed and trained to identify COVID-19. This model was developed using a deep learning algorithm to detect patients with COVID-19 using chest x-rays. The model was trained and tested on extensive dataset from US healthcare system. The study compares the result of the model with the interpretations given by experience thoracic radiologists.

The key areas of focus were on:

1. Image labelling and Dataset Partitioning: Regardless of the quality of the image, data were gathered if the patients met the inclusion criteria for the study.
2. Ensemble of DeepCOVID-XR: 24 individual neural network architectures were used to build the model. Details like image pre-processing, algorithm training, validation and testing were performed.
3. Experienced thoracic radiologist's interpretations: Randomly 300 chest x-rays were selected for 5 radiologists to provide interpretations on. Radiologists were not provided with any identifiers or clinical information and were given access to PACS. They provided their interpretations on radiographs as being positive or negative for COVID-19 with a confidence level as -3 being highest level for negative and +3 for positive. This 6 scoring system was calculated at 5 various thresholds for each radiologists to compare it with the model predictions. A consensus interpretation was evaluated by taking the majority vote of the single radiologist's interpretations and a ROC curve was plotted by calculating the average of 6 point scores for all 5 radiologists on every image.
4. Performance of the Algorithm: After performing the validation out of 2214 images, DeepCOVID-XR had an accuracy of 83% and yielded sensitivity of 75% with specificity of 93%. The AUC was 0.88 that was plotted on the ROC graph that was compared with AUC of consensus interpretation of 5 radiologists which was 0.85.

   When algorithm was compared with the interpretation of radiologists as a standard reference it yielded an AUC of 0.95 rather than RT-PCR assay

2.2 Literature Review

Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers

The study was conducted to weigh out the design characteristics of the included published studies that reports of AI algorithm's performances that interpret medical images and regulate if the study designs were appropriate for justifying the performance of AI algorithms.

The methodology that was adopted for this study was to search and include original research papers between a specified timeline (January 1, 2018 and August 12, 2018) that had validated the performance of AI algorithms to cater diagnostic determination. The inclusion criteria for this study was: which validation was conducted external or internal, if external, collection of validation data was prepared, the study design chosen for all validations.

The results of this study concluded that out of 516 acceptable articles on 31 studies, that comprises of 6% of external validation. Also, none of the 31 studies followed all three design features**.**

Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction

This study was done to illustrate primary methodological point's involved validating AI technology for its usage in medicine, more importantly diagnostic and predictive models for which deep learning methods are used. This paper also includes the effects of disease manifestation spectrum and disease prevalence on the performance results which is in continuation with discussing the difference between evaluating the performance with use of internal and external datasets.

At the end authors comment on the role of trials and outcome studies for their clinical verification of diagnostic or predictive AI tools through patient outcomes, beyond performance metrics and also how these kind of studies can be designed.

The Algorithmic Audit: Working with Vendors to Validate Radiology-AI Algorithms - How We Do It

With plenty of AI tools being advanced around the world which aspire to either boost up or enhance the certainty of clinical expertise. For developers and radiologists it is imperative to work together to determine accurate clinical utility and liability associated with these models. This paper showcase a plan to work with such developers that builds these AI tools to assess and improve the performance of these models. The plan comprises of concepts of accurate autonomous evaluation on data that the model has not seen previously, curating data for such validation, profound examination of false positives and false negatives, also to audit the indication of such flaws and real-time deployment and validation of AI models.

Testing and Quality Validation for AI Software– Perspectives, Issues, and Practices

This paper spotlights on quality testing for AI software functionality features. It provides an insight for new features and requirements. In inclusion to this other testing classes and ways are showcased. Also, illustration of quality assessment and criteria are presented.
Furthermore, a practical study on quality validation for an image recognition system is performed through a meta-morphic testing method. With new AI tool features, challenges and concerns for testing software and quality of the system have emerged.
This study also discusses the existing validating approaches and their analysis, with evaluation of validation quality and reporting problems in AI software.

DeepCOVID-XR: An Artificial Intelligence Algorithm to Detect COVID-19 on Chest Radiographs Trained and Tested on a Large U.S. Clinical Data Set

This paper reports about an AI algorithm to detect COVID019 on chest images. The algorithm was developed and validated on considerable amount of dataset, DeepCOVID-XR.

The model was developed to detect COVID-19 on chest x-rays that was developed and validated on an extensive dataset. The model was tested on 14788 images out of which 4253 were COVID-19 positive, taken from different locations from February 2020-April 2020 and then those were validated on 2214 images out of which 1192 were positive for COVID-19. Ground Truth was prepared by five experienced thoracic radiologists, 300 random test images For all the dataset, the accuracy of the algorithm was 83%, with AUC of 0.90. The AUC was 0.88 that was plotted on the ROC graph that was compared with AUC of consensus interpretation of 5 radiologists which was 0.85.When algorithm was compared with the interpretation of radiologists as a standard reference it yielded an AUC of 0.95 rather than RT-PCR assay.

2.3 Methodology

The study was carried out at CARING department of Mahajan Imaging Research and Development Pvt. Ltd. It is a descriptive study design with secondary research.

This study includes 3 case studies which were performed from 1st March-1st June using CARPL platform as a tool for validating various AI models on different types of medical images.

Research articles were extracted using PubMed, google scholar and various other databases that included research studies on validation of AI models that investigate diagnostic decisions/prediction of medical images.

Inclusion Criteria:

1. Research studies on validation of AI models
2. Challenges faced during performance testing of AI models
3. Computing ways to improve AI algorithms

Exclusion Criteria:

Articles that did not involve evaluation of performance of AI algorithms on medical images.

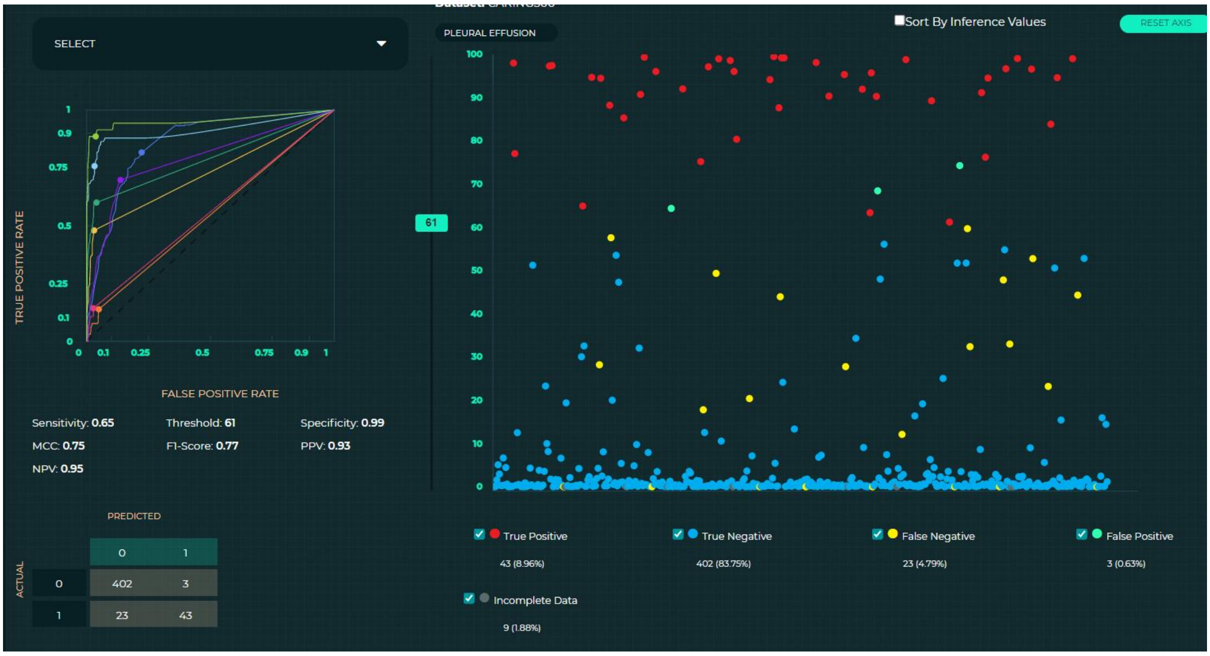
Three Validations were conducted for three different AI models using CARPL platform as a statistical tool.

2.4 Case Study

1. Conducted a Validation for an AI model on Chest X-rays for various Lung Abnormalities

Process:

a. First and foremost you need a dataset to conduct a validation of an AI model. I chose the standard dataset available on CARPL which contains 500 chest x-rays for the validation.
b. Secondly, as for validation we need inference results from the algorithm and ground truth to conduct a pre-deployment testing. Using the platform I uploaded the inference results which I obtained from the algorithm and updated the ground truth.
c. Lastly, I validated the performance of the AI model, the image depicts ROC curve on the left hand side and shows an AUC value for one of the various abnormalities AI model was trained and developed for. On right hand side, the dots predicts the images that were used for the validation. Different colors depicts different kinds of results.
   1. Blue-True negatives
   2. Red-True positives
   3. Green-False positives
   4. Yellow-False negatives



Performance of the AI model:

1. AUC for Pleural Effusion is 0.911, and as explained closer the AUC to 1, better is the performance of the model. Here the model gives the probability of approx. 91% that it can detect images with Pleural Effusion correctly.
2. ROC curve changes with change in the threshold. To study the performance of an algorithm dynamic thresholds are important to locate false positives and negatives so that model can be trained accordingly.
3. Here on the above image, the threshold has been set at 61, as the model best performs at that threshold by giving F1 score of 0.77 and specificity of 0.99.

2. Conducted a Validation for an AI model on Chest X-rays for various Lung Abnormalities

Performance of the AI model:

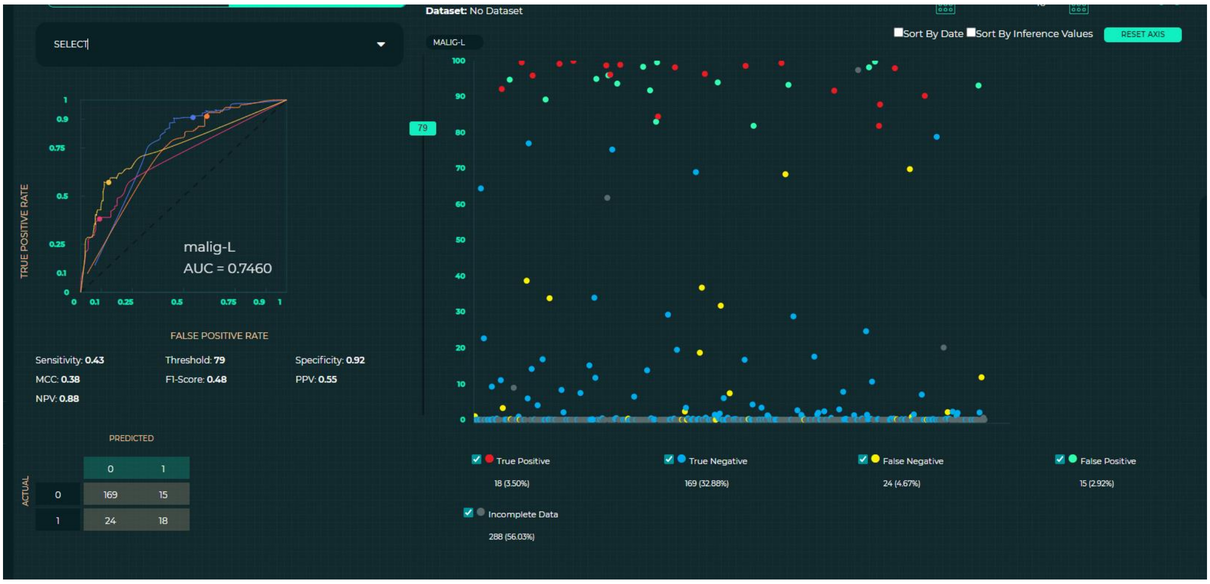
1. This model gives an AUC value of 0.975 for Pleural Effusion. This means the model can detect images of Pleural Effusion with a probability of 97%.
2. This algorithm best performs for this abnormality at a threshold of 71, giving F1 score of 0.83 and specificity of 0.94 and sensitivity of 0.96

If you compare the two models only for Pleural Effusion, then it clearly states that AI model 2 performs better and has greater probability of detecting the abnormality correctly.

3. Conducted a Validation for an AI model on Mammograms for BIRADS Classification

Process:

a. 514 Mammograms were used as testing dataset for this validation.
b. Secondly, as the platform provides an output in a binary classification the BIRADS classification were divided into Malignant and Non-Malignant cases.
1. BIRADS 1, 2 and 3 were considered Non-Malignant
2. BIRADS 4, 5 and 6 were considered Malignant
c. The inference results and ground truth were uploaded as for Malignancy and Non-Malignancy for both the breasts and performance of the model was studied.



	0	1
0	169	15
1	24	18

Performance of the AI model:

1. The Mammo Algorithm gives an AUC value of 0.7460 for Left breast Malignancy. That means the algorithm has the probability of 74% of correctly identifying Malignancy in the left breast images.
2. When threshold set at 79, the model yields better statistics with F1 score of 0.48 and specificity of 0.92.

2.5 Results

To successfully validate an AI model/algorithm few methodologies that should be considered are:

1. Validation Imperatives: The two fundamental factors for conducting a validation are Datasets and clinical proficiency. To evaluate an AI algorithm an extensive amount of dataset is required which are clinically significant in order to outstand as how the algorithm performs in different situations. In addition to large datasets, clinical judgment is also an important factor as it helps to understand about the performance of an algorithm. In many situations it is difficult to evaluate why an algorithm failed. In such instances help of a clinical expertise is compelling on analyzing failed cases and provide reasons for it.

2. Absolute Validation: For any AI algorithm to be used in a clinical domain, its generalizability is critical and also a challenge. During the development of an algorithm a dataset is used which is known as training dataset that is used for fine-tuning the parameters of the algorithm. Addition to training dataset, a testing/validation dataset is required for evaluating the performance of the algorithm. If an external dataset is used for validation it is known as external validation of the algorithm. Since the model would perform better on internal dataset, an external validation is critical for determining the performance the algorithm before deploying it in a clinical practice. If an algorithm is developed, trained and tested from using the data of one site it may not provide the accurate results of the performance of the AI model.

3. Selection of the Data mix: Once the ground truth is prepared it is critical to evaluate the mix of cases to validate the AI model. Generally AI model provides two such type of failures, one being False Positives, where the model gave an output of the image having the abnormality but in reality it was negative, second is False Negative where AI gave an output for the image being normal instead it was abnormal in reality. Now to study the FP rate, a dataset without many positives are required as to evaluate how often did the model missed negative cases and stated them as positives. Similarly, to study the rate of FN, a dataset without many negative cases are required, to determine how frequently the algorithm did missed actual positives cases and stated them as negatives.

Techniques for improving an AI model:

1. Auditing of FN and FP: To improve the algorithm, an extremely important method is auditing of false negatives and positives. To do this a data scientist and radiologist has to work together and look for cases where the model failed. Indications of where AI went wrong should be studied aggressively, for example a model missed detecting broncho-pulmonary markings is less risky than AI missing large pneumothorax. Another way is to provide reasons for model failure. For every FN and FP cases a reason should be provided why the model failed here. This would help the developers of the model to look for a pattern and improve the algorithm.

2. Dynamic Thresholds: This is another way to improve the output of the AI model. Dynamic threshold means a value to evaluate whether an abnormality is present or not which changes with change in the clinical situations. This reduces the errors caused by the model for example, a patient who came for a routine health check-up got a chest x-ray, for this patient the threshold for detecting any lung abnormality should be high, whereas compared to a patient who is in ICU. Addition of clinical significance would decrease the errors and improve the output by the model.

2.6 Conclusion

- In the world of AI, the term validation refers to fine-tuning the parameters for algorithm development and the test is used to evaluate the algorithm performance.
- There are two aspects to evaluate the performance of any predictive model i.e Discrimination and Calibration which are usually determined by plotting ROC curve and Calibration curve respectively.
- Another critical or one of the essentials for any validation of an AI model is external validation which increases the robustness of the algorithm. An extensive amount of data should be considered for evaluation of any model.
- Lastly, how the model would benefit patients in long run should be a point of concern, which can be avoided by performing number of clinical trials on the algorithm before deploying it on a clinical practice.

2.7 References

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6389801/

2. https://pubs.rsna.org/doi/10.1148/radiol.2017171920

3. https://vixra.org/pdf/1909.0104v1.pdf

4. https://www.sciencedirect.com/science/article/pii/S2666389920300428

5. https://www.researchgate.net/publication/335371124_Testing_and_Quality_Validation_for_AI_Software-Perspectives_Issues_and_Practices

6. https://pubs.rsna.org/doi/full/10.1148/radiol.2020203511